

AD \_\_\_\_\_

Award Number: DAMD17-98-1-8323

TITLE: Deriving Structures for Lead Drug Discovery from Cell-Line Screens

PRINCIPAL INVESTIGATOR: Doctor Robert L. Jernigan  
Doctor David G. Covell

CONTRACTING ORGANIZATION: National Cancer Institute  
Bethesda, Maryland 20892

REPORT DATE: October 2001

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command  
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;  
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

20020909 024

**REPORT DOCUMENTATION PAGE**Form Approved  
OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

<b>1. AGENCY USE ONLY (Leave blank)</b>		<b>2. REPORT DATE</b> October 2001	<b>3. REPORT TYPE AND DATES COVERED</b> Final (1 Sep 98 - 1 Sep 01)	
<b>4. TITLE AND SUBTITLE</b> Deriving Structures for Lead Drug Discovery from Cell-Line Screens			<b>5. FUNDING NUMBERS</b> DAMD17-98-1-8323	
<b>6. AUTHOR(S)</b> Doctor Robert L. Jernigan Doctor David G. Covell				
<b>7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)</b> National Cancer Institute Bethesda, Maryland 20892  E-Mail: <a href="mailto:covell@helix.nih.gov">covell@helix.nih.gov</a>			<b>8. PERFORMING ORGANIZATION REPORT NUMBER</b>	
<b>9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)</b> U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012			<b>10. SPONSORING / MONITORING AGENCY REPORT NUMBER</b>	
<b>11. SUPPLEMENTARY NOTES</b> Report contains color				
<b>12a. DISTRIBUTION / AVAILABILITY STATEMENT</b> Approved for Public Release; Distribution Unlimited				<b>12b. DISTRIBUTION CODE</b>
<b>13. ABSTRACT (Maximum 200 Words)</b> <p>The primary aim of this research was to develop a suite of computational tools for the examination of tumor screening data from the NCI's tumor screening databases. These tools were designed to easily process data from over one hundred immortalized tumor cells screened for growth inhibition by over 30,000 synthetic compounds. This analysis consisted of a self-organizing-map (SOM) clustering of compounds based on their screening responses. Our results find that clearly defined classes of compounds are clustered based on their mechanism of action. Six general groupings were identified according to the broadly defined putative classes of cellular action for these agents: nucleic acid biosynthesis, mitosis, kinase and phosphatase signaling pathways, membrane function (integrity and transport), protein metabolism, and a class of agents that exclude the previous 5 classes, and have not yet been associated with a particular cellular function. These results provide a facile means of relating previously screened compounds to the large libraries of untested compounds. This effort will increase opportunities for the discovery of novel anti-tumor agents.</p>				
<b>14. SUBJECT TERMS</b> Tumor Screening, Bioinformatics, Statistics				<b>15. NUMBER OF PAGES</b> 84
				<b>16. PRICE CODE</b>
<b>17. SECURITY CLASSIFICATION OF REPORT</b> Unclassified	<b>18. SECURITY CLASSIFICATION OF THIS PAGE</b> Unclassified	<b>19. SECURITY CLASSIFICATION OF ABSTRACT</b> Unclassified	<b>20. LIMITATION OF ABSTRACT</b> Unlimited	

## Table of Contents

<b>Cover.....</b>	<b>1</b>
<b>SF 298.....</b>	<b>2</b>
<b>Introduction.....</b>	<b>4</b>
<b>Body.....</b>	<b>5</b>
<b>Key Research Accomplishments.....</b>	<b>6</b>
<b>Reportable Outcomes.....</b>	<b>6</b>
<b>Conclusions.....</b>	<b>7</b>
<b>References.....</b>	<b>-</b>
<b>Appendices.....</b>	<b>8</b>

## **Introduction:**

**The primary aim of this research was to develop a suite of computational tools for the examination of tumor screening data from the NCI's tumor screening databases. These tools were designed to easily process data from over one hundred immortalized tumor cells screened for growth inhibition by over 30,000 synthetic compounds. This analysis consisted of a self-organizing-map (SOM) clustering of compounds based on their screening responses. Our results find that clearly defined classes of compounds are clustered based on their mechanism of action. Six general groupings were identified according to the broadly defined putative classes of cellular action for these agents: nucleic acid biosynthesis, mitosis, kinase and phosphatase signaling pathways, membrane function (integrity and transport), protein metabolism, and a set of agents unassigned to these five classifications, that represent future efforts to identify their cellular function. These results provide a facile means of relating previously screened compounds to large libraries of untested compounds. This effort will increase opportunities for the discovery of novel anti-tumor agents**



**Body:**

The research completed over the funding period closely followed the objectives provided in the statement of work(SOW). The initial aims were to develop computational tools for the analysis of large biological databases relevant to the discovery of agents with possible anticancer activity. Much of the initial effort was devoted to a systematic and thorough analysis of the NCI's publicly available tumor screening databases. This initial effort was equally divided into a) the development of database-related navigational tools for data stratification and visualization, and b) the development of computational means for assessing the information content of these databases. These efforts revealed the need for additional computational tools for data examinations and detailed considerations of noisy data versus meaningful information. All of this effort is summarized in the attached publication (see Appendix). Two noteworthy points are appropriate in reference to this publication: its considerable length and the managing editor's willingness to accept a manuscript of this size. The former feature is appropriate for the breadth of detail necessary for a systematic and comprehensive analysis of this data. The latter feature was based on the strong supporting information provided by the reviewers' of this manuscript for this research effort.

Following the recommended instructions for this report, the accepted manuscript is included in the appendix.

## **Key Research Accomplishments:**

- **The first comprehensive analysis of the NCI's tumor screening databases**  
**Over 36K synthetic compounds screened against over 100 tumor cells**
- **The systematic development of methods for assessing quality of data**  
**Noise versus Signal**  
**Effect of missing data**  
**Effect of partial screens against sub-optimal tumor cell numbers**  
**Contrasts with alternative statistical analyses**
- **Complete SOM organization of biological response for 36K compounds**
- **Identification of six broadly defined classes of cellular activity**  
**Nucleic acid biosynthesis**  
**Mitosis**  
**Cellular Signaling**  
**Membrane integrity and transport (apoptosis active agents)**  
**Protein Metabolism Poisons**  
**Agents not in the above classes**
- **Development of a web-based tools (<http://spheroid.ncifcrf.gov>) for a wide range of solicitations of this data**  
**Compound searching and sub-searching**  
**Access to publications related to screened compounds or their mechanism of action**  
**Access to other databases (pdb, pubmed, unigene, omim)**  
**Visualization of raw data vectors of cellular growth inhibition**  
**Visualization of structural information**  
**Tools for hypothesis generation**

## **Reportable Outcomes:**

- **Manuscript to appear in The Journal of Medicinal Chemistry.**  
**Title: Mining the NCI's Tumor Screening Database: Identification of Compounds with Similar Cellular Activities**
- **Presentation at National Meetings:**  
**Protein Society, 2000 and 2001**  
**Biophysical Society, 2000 and 2001**  
**NIH Division Of Basic Science Annual Retreat, 1999-2001**  
**Era of Hope Meeting, Atlanta, 2000**

## **Conclusions:**

**Our research effort represents the first comprehensive analysis of the complete set of compounds available in the NCI's database of publicly screened compounds. The novelty of this approach has yielded a unique perspective to the classes of agents currently available for cancer therapy as well as an assessment of alternative compounds with potentially comparable cellular activities. These computational tools have proven quite effective for the identification of putative mechanism of action for these compounds and the formulation of hypotheses regarding chemical modalities responsible for cellular activities. Noteworthy amongst our research is the development of a web-accessible tool ([spheroid.ncifcrf.gov](http://spheroid.ncifcrf.gov)) that provides public access to this data as well as our published results.**

**Appendix:**

Mining the NCI's Tumor Screening Database: Identification of  
Compounds with Similar Cellular Activities

Alfred A. Rabow<sup>1</sup>, Robert H. Shoemaker<sup>2</sup>,  
Edward A. Sausville<sup>2</sup> and David G. Covell<sup>2</sup>

<sup>1</sup>Science Applications International Corporation,

<sup>2</sup>Developmental Therapeutics Program, DCTD,

National Cancer Institute, NIH

Frederick MD 21702, USA

November 15, 2001

Corresponding Authors:

Alfred A. Rabow and David G. Covell

Laboratory of Computational Technologies, Screening Technologies Branch,  
DTP, NCI, NIH, SAIC

Bldg. 430, Room 215

NCI-Frederick, Frederick, MD 21702

301-846-5785 (voice) 301-846-5762 (fax)

[rabowa@ncifcrf.gov](mailto:rabowa@ncifcrf.gov)

[covell@ncifcrf.gov](mailto:covell@ncifcrf.gov)

# 1 Abstract

In an effort to enhance access to information available in the NCI's anti-cancer drug screening database, a new suite of internet accessible (<http://spheroid.ncicrf.gov>) computational tools has been assembled for self-organizing-map-based (SOM) cluster analysis and data visualization. A range of analysis questions were initially addressed to evaluate improvements in SOM cluster quality based on the data conditioning procedures of Z-score normalization, capping and treatment of missing data, as well as completeness of drug cell screening data. These studies established a foundation for SOM cluster analysis of the complete set of NCI's publicly available anti-tumor drug screening data. This analysis identified relationships between chemotypes of screened agents and their effect on four major classes of cellular activities: mitosis(M), nucleic acid synthesis(S), membrane transport and integrity(N) and phosphatase and kinase mediated cell cycle regulation(P). Validations of these cellular activities, obtained from literature sources, found i) strong evidence supporting within cluster memberships and shared cellular activity, ii) indications of compound selectivity between various types of cellular activity and iii) strengths and weaknesses of the NCI's anti-tumor drug screen data for assigning compounds to these classes of cellular activity. Subsequent analyses of averaged responses within these tumor panel types finds a strong dependence on chemotype for coherence among cellular response patterns. The advantages of a global analysis of the complete screening data set are discussed.

## 2 INTRODUCTION

The vast amounts of data accompanying the post-genomic era have raised many exciting questions about new experimental designs and methods of data analysis [1, 2, 3, 4]. High-throughput biological assays [5, 6], methods of chemical synthesis [7, 5] and genome analysis of protein and gene expression arrays [8, 9, 10] have spawned this explosion of biotechnology data. The information contained in many of these experiments lies in the diversity of measurements across different test systems and the utilization of this information as probes into underlying biological phenomena [11, 12, 13]. Systematic investigations of these patterns offer the promise of facilitating new drug discoveries and improving the molecular taxonomy for chronic diseases such as cancer and associated opportunistic infections [14, 15, 16, 17].

Applications of computer-supported analytical techniques combined with interactive and dynamic data visualization tools are needed to assist in the goals of discovery, decision-making and explanation [18]. From research in statistical methods that began in the early 1960's, a wide-range of tools are now available for the analysis of multi-dimensional data that includes techniques of hierarchical clustering [19], k-means clustering [20, 21], multidimensional scaling [22], binary deterministic-annealing [23] support vector machines [24] and self-organizing maps [25, 26]; each method aimed at the identification of pattern similarities between diversity measures. While each of these methods has a sound foundation in mathematical statistics, their applications to large biological data sets can be difficult, vary in their suitability to each design problem, and often lack required assessments in the critical areas of reliability, post-analysis validations, internal consistency and data conditioning, among other analytical issues. Of additional importance with respect to the final product (i.e. the clustering), these results should be readily accessible and provide facile understanding to the intended audience, e.g. the drug discovery and biochemical communities.

This work describes a new set of computational tools, based on methods of self-organizing maps (SOMs), and their applications to the NCI's anti-tumor drug screening data. This suite of tools seeks to identify compounds with similar activities against these tumor cell lines and thereby facilitate discoveries of potentially new drug leads and new molecular targets. The report contained herein is divided into two parts. The first part addresses analytical questions

related to data analysis. A number of areas are presented that affect cluster quality, including: (1) the number of cell lines in the screen; (2) the size of the SOM clustering map; (3) the treatment of noisy and incomplete data; and, (4) the importance of data conditioning. The results of these studies establish a foundation for the second part of this report which documents our analysis of the complete set of the NCI's publicly available drug screening data. This second section is focused primarily on critical examinations of cluster memberships and validation of their putative cellular activity in this screen. These results provide a classification scheme for relating the activity of groups of compounds to cellular processes or, when possible, to specific molecular targets. Graphical tools have been constructed that serve to aid in data visualization and data analysis. The suite of tools comprising our approach has been assembled into an internet accessible methodology (<http://spheroid.ncifcrf.gov>), referred to as the 3d MIND (Mining Information for Novel Discoveries) toolkit. While this presentation is focused on anti-tumor drug screening data, these methods are general and should find applicability in the analysis of data sets based on a variety of biologically diverse measurements, including the new generation of microarray data sets and data from molecular targeted high-throughput screens.

## 2.1 Screening Data

Conceived in the late 1980's and implemented in the early 1990's, the NCI has assembled an extensive screening database of anticancer compounds and diverse chemicals of unknown biological activity tested against tumor cells [27, 28, 29, 30]. This database (<http://www.dtp.nih.gov>) contains measures of growth inhibition ( $[\log(GI_{50})]$ ) total growth inhibition (TGI) and fifty percent cell killing (LC50) for over 100,000 compounds tested in various subsets of 60-100 tumor cell lines [28]. While all of these measures reflect a variety of biological processes involved in cellular proliferation, the results reported here are focused on an analysis of the  $\log(GI_{50})$  data.

Computer-based tools have been developed that analyze *patterns* of  $\log(GI_{50})$  measurements against the tumor cells within the screen [27, 3]. Based principally on the analysis of pairwise statistical correlations between drugs tested against various tumor cell lines, similarities in mechanisms of action, modes of resistance and molecular structure have been re-



vealed [31, 32, 33, 34]. While these methods have provided very useful information for relating molecular structure to its putative biological function [35], they are somewhat cumbersome, producing lists of compounds whose structural and biological features require extensive manual analysis. Quantification of the improvements in data analysis resulting from our 3d MIND clustering methodology is described in the appendix.

Motivated by the premise that alternative methods of data analysis may be useful for extracting additional information from the NCI's tumor screening database, we have developed a suite of tools, based on self-organizing maps (SOMs), useful for data exploration and especially, its biochemical interpretation via two-dimensional biological response maps. This suite of tools is focused in the areas of data conditioning, pattern association, visualization and data presentation, with additional functionalities that address signal scaling issues, missing data elements, and locality/non-linearity features of the data space. As we will demonstrate herein, careful and critical considerations in these areas can enhance the extraction of additional information from large, complicated, screening databases as well as provide a general tool well suited for drug discovery investigations.

### 3 METHODS

While no standard method exists for the analysis of large and complex data sets, a detailed understanding of the experimental design and methods of analysis is essential for evaluating the results of data explorations. With this goal in mind, this section will present a brief description of the screening experiments, the conditioning of data, and the statistical methods involved in data analysis and display. This information will provide a rationale for later design and data analysis decisions.

#### 3.1 Data Conditioning

The NCI's anti-cancer drug cell screen generates measures of the fifty percent growth inhibition,  $\log(GI_{50})$ , of selected established tumor cell lines following exposure to test compounds [29, 30]. While the 'raw' data generated in this screen determines a compound's potency for growth

inhibition, most of the interest in this data lies in establishing the biological significance of the cellular response pattern; with the hope of identifying tumor selective reagents, new molecular targets and new drug lead compounds. We treat this raw data with Z-score normalization to enhance the biological response signal.

Z-score conditioning of each compound’s raw data against the panels of tumor cells provides a common mean reference and scale thus enhancing the cellular response signal:

$$z_{ij} = \frac{(g_{ij} - \bar{g}_j)}{\sigma_j}, \quad (1)$$

where  $z_{ij}$  is the z-score for compound  $i$ , against tumor cell  $j$ ;  $g_{ij}$  is the measured  $\log(GI_{50})$  value; and,  $\bar{g}_j$  is the mean and  $\sigma_j$  is the absolute deviation across all cell types  $j$ . Using a metric related to data clustering (to be presented later) we find that the Z-score transformed data improves the quality of the clustering by  $\sim 15\%$  when compared to the raw data. An additional consideration for data conditioning involves the intrinsic sensitivity of cell lines to chemical agents. For example, analysis of the  $\log(GI_{50})$  values for the NCI synthetic compound data set finds the leukemia (LEU) cell panel is most sensitive to chemical agents, whereas, the non-small cell lung (LNS) panel is the least sensitive. Data normalization facilitates assessment of the differential growth inhibition across all cell lines, rather than detecting agents active against only the most sensitive cell lines. Z-score normalization of each cell line’s response to all tested compounds thus establishes a common reference. Alternatively, scaling the raw data across tumor cell types *and* across tested compounds provides a uniform means to assess pattern diversity within the complete set of tumor screening data. <sup>1</sup>

### 3.2 Self-Organizing Maps

Traditional methods for mining large screening data sets seek to discover subsets of data where similarities in response are observed. The initial step in this process requires the selection of a pairwise measure of pattern similarity that assigns the highest score to the most similar

---

<sup>1</sup>Data from all tumor cell lines was used in our analysis. This set consists of 80 cell-lines collected from leukemia(LEU), non-small cell lung(LNS), small cell lung(SCL), colon(COL), central nervous system(CNS), melanoma(MEL), ovarian(OVA), renal(REN), prostate(PRO) and breast(BRE) cancer tissues.

data sets. Such pairwise measures include rank correlation and Euclidean, Mahalanobis or Minkowski measures of distance [19]. These pairwise measures provide a simple and direct means to identify highly similar response subsets. Limitations in this procedure are known to occur, particularly when data is contaminated with large amounts of noise, resulting in a greater likelihood of random statistical correlations, and increased difficulty in determining 'real' relationships [19, 36]; a result particularly evident where data cannot be reduced by simple bisection into groups (i.e. pairwise hierarchical clustering). Methods designed to treat noisy data include principal component analysis and the related method of singular value decomposition; where the data are re-expressed along directions that maximize the signal-to-noise ratio [35]. The self-organizing map (SOM) method [25] has found great utility in studies of voice recognition and visual processing; data sets which often exhibit large amounts of random noise and missing data [37, 36]. Designed specifically to deal with extremely noisy and incomplete data sets, the algorithms associated with the SOM method are well-suited for mining data from the NCI's anti-cancer drug screen.

The SOM [38] method can be divided into two regimes: clustering in high dimensional space, and projections into a lower dimensional display space (see Figure 1). This first step clusters data in its original high dimension space (for the NCI screen  $N=80$ .) The SOM algorithm locates the response vectors in this high dimensional data space by minimizing the deviation between the data vectors( $V^j$ ) and response vectors( $R^k$ ): *Fig. 1.*

$$\nabla R^k \propto \sum_j h(||V^j - R^k||) ||V^j - R^k|| \quad (2)$$

where  $\nabla R^k$  is the incremental change in position of the response vector  $R^k$ ,  $V^j$  is the set of data vectors, and  $||V^j - R^k||$  is the distance between data and response vector. The neighborhood kernel function,  $h(||V^j - R^k||)$ , weights the change in the position of the response vectors. This neighborhood kernel collectively orders the response vectors to mirror the information contained in the data space [25]. The form of the neighborhood kernel function exhibits a maximum when the data and response vectors coincide and goes to zero as these vectors become more distant. Often the neighborhood kernel is a Gaussian function, however, our analysis finds that Epanechnikov function  $[\max(0, 1 - ||V^j - R^k||^2)]$  consistently yields improved clustering, and was used for our analysis.

The form of equation 2 determines the position of response vectors that best matches the data space, or alternatively, how the response vectors partition the data space into clusters (see Figure 1, Panels B and C). Regions that are rich in data vectors attract many response vectors and as a result finely divide these regions of high information content. This process can be contrasted with the more conventional principal component analyses, where data is oftentimes reoriented, in a linear fashion, on to the space of the top most principal components. The biochemically important regions of the NCI's anti-cancer drug screening data are not uniformly distributed in the 80 dimensional tumor cell space, but rather are contained in densely populated sub-spaces. This distribution can be quantified by examining the pair-wise Euclidean distance distribution. The mean pair distance of the normalized  $\log(GI_{50})$  data found to be  $10.34 \pm 1.57$  deviation units compared with the value found with a uniform distribution  $21.85 \pm 1.46$ . The SOM transformation stretches these data rich regions, thereby enhancing biochemically relevant cluster distinctions and matching the underlying data distribution (mean pair distance of the SOM coordinates is  $10.18 \pm 1.48$ ). A direct consequence of SOM reordered data is the ability to display these results in an interpretable manner. The method of display is the uniform projection of SOM clustering in high-dimensional space to a low dimension display space (see Figure 1, Panel D). This mapping is both simple and retains a great deal of the original high-dimensional information. Additional details regarding the application of the SOM method to the NCI's tumor cell screen can be found in the 'Overview' and 'Tutorial' sections of the 3d Mind web pages (<http://spheroid.ncifcrf.gov>.)

## 4 RESULTS

We begin with an analysis of tumor growth inhibition by a set of 122 standard anticancer agents (<http://www.dtp.nci.nih>) compiled by Weinstein et al. and annotated according to their putative mechanism of action (MOA) [4, 39]. Figure 1, Panel D, displays the two-dimensional SOM map for an extended data set comprised of compounds structurally similar to these standard agents. Consistent with prior studies, these standard agents could be separated into those with MOA's involving inhibition of mitotic activity and those affecting nucleic acid biosynthesis [35]. This division is quite sharp, and appears in Figure 1, Panel D, below row

six of the SOM map. Within these two regions of the map, well defined sub-clusters exist that, upon inspection, consist of structurally similar compounds with stick-figure drawings of selected cluster members displayed at the map margins. This apparent consistency between molecular structure and function (putative MOA) was used to develop a metric for detailed sensitivity studies regarding the choice of parameters for later SOM optimizations and their effect on quality of clustering.

## 4.1 Sensitivity Analysis

This analysis attempts to determine the relationship between quality of clustering and choices in the experimental design parameters of number of cell lines in the screen, size of the SOM clustering map, treatment of noisy and incomplete data, and importance of data conditioning. We assess the quality of clustering by correlating the SOM cluster memberships determined from the  $\log(GI_{50})$  (i.e. functional) data with the SOM clustering based on chemical structure (i.e. structural). This approach assumes that chemical structure, as defined by atom type and bond connectivity, is a surrogate for the 'true' pharmacophore of the molecular target affecting cell growth. This is clearly a simplifying approximation for the true 'hidden' pharmacophore or molecular target [41]. Indeed, small structural modifications are known, in some cases, to radically alter biological activity.

To examine the correlation between cluster memberships based on biological response and chemical structure, we have designed an extended mechanism of action (ExMOA) data set which consists of 362 compounds, based on the original set of 122 standard anticancer agents discussed above, but expanded to include screened compounds with strong structural similarity (Tanimoto coefficient  $> 0.9$ ) [42, 43, 44, 45] to these standard anticancer agents. SOM clustering of these compounds into structural classes is based on the E-state bit vectors available in the CACTVS suite of computational tools (<http://www2.chemie.uni-erlangen.de/software/cactvs>). These bit vector assignments represent 431 chemical descriptors developed within CACTVS, with characteristics similar to assignments available within the MDL ISIS keys [46]. SOM clustering treats the vectors of 431 structural descriptors for each agent in the same fashion as the vectors of  $\log(GI_{50})$  values used for SOM clustering of the biological data. The correlation

between biological clusters and structural clusters was accomplished with an heuristic matching algorithm. This approach uses dynamic programming to order the structural and functional clusters to achieve the greatest fractional overlap of cluster members. As an example, consider nine compounds, labeled *a* through *i*, that are mapped to three clusters according to their structural bit vectors; [a,b,c] [d,e,f] [g,h,i], and three clusters according to their biological response; [b,e,f,g] [a,c] [d,h,i]. Based on a heuristic of maximal cluster overlap, and using as a reference the structural cluster order, the functional clusters would be reordered as [a,c] [b,e,g,f] [d,h,i], where the underlined letters indicate shared elements between structural and functional clusters. The measure of cluster quality is determined as the linear correlation coefficient between these reordered lists of compounds. Although this is a rather general measure of cluster quality, it correctly reflects the overlap of individual cluster memberships when clustering is achieved with two different methods, i.e. one based on structural descriptors and the other based on cellular response. Example structure/function (S/F) plots are shown in Figure 2. It should be noted that what is chiefly of interest is the change in S/F correlation, not the absolute quantity. Therefore, any measure that accurately reflects relative correlation will serve as a surrogate marker for quality in the sensitivity analysis. Alternative assignments of structural bit vectors (i.e. MDL keys) or using the biological clusters as the 'reference' order does not significantly alter the following results. Fig. 2.

Table I lists the correlations between cluster memberships determined from biological response data [ $\log(GI_{50})$ ] and chemical structure (bit vectors) for different data conditioning treatments. We have found a 15%  $[(0.9002-0.7820)/0.7820]$  improvement in the correlation coefficient with the Z-score normalization over an analysis based on raw data. This improvement is statistically significant, with an ANOV1 p value of  $1.7e-15$ ; a clear indication that Z-score normalization enhances the quality of clustering. In addition to Z-score normalization, the magnitude of any component of a data vector has been capped at a value of  $\pm 3$  absolute deviation units from the vector mean. Capping prevents the difference between two data vectors being dominated by a single or a few cell lines which have extreme values. Avoiding strong outliers by data capping improves the S/F correlation by 2.0%  $[(0.9185-0.9002)/0.9002]$ . This apparently small improvement is, however, statistically significant, with an ANOV1 p value of  $4.4e-6$ , and has been adopted as a feature of data conditioning. Tbl. I

Another important design choice addresses the treatment of missing data. Oftentimes missing data are replaced by their mean value determined from existing data. Our analysis indicates that this approach can substantially distort the information contained within the actual data. Retaining missing data elements as unknowns, rather than replacement by their vector mean, improves the S/F correlation coefficient by 6%  $[(0.9185-0.8654)/0.8654]$ . This improvement is statistically significant, with an ANOV1 p value of 7.6e-14, and supports our choice to treat missing data as unknown. Figure 2, Panels A, B and C, display the S/F correlations for selected cases of data conditioning. Panels A and B represent the best and typical S/F correlations, respectively, for applications of Z-score normalization, capping and no substitutions for unknowns(NaN). Panel C is an example where a poorer S/F correlation results when unknown data are replaced by their group mean.

## 4.2 Map Dimensions

The possibility that map dimensions may affect the quality of the clustering was investigated using S/F correlations. The SOM method contains a heuristic for the ratio of the two-dimensional SOM map dimensions based on the ratio of the two largest eigenvalues as the linear SVD solution to the data set [25]. Using this heuristic and the ExMOA data set, the SOM analysis recommends a map size of 17x9, based on an eigenvalue ratio of 1.89. Figure 3 displays *Fig. 3.* the dependence of the S/F correlation coefficient for a selection of map ratios. The ratio that maximized the correlation coefficient matched the heuristic at 1.89. Ratios above and below this value generate maps with a concomitant degradation in their S/F correlation. In connection with the non-square map ratio of 1.89, the clustering algorithm does not use "wrap-around" boundary conditions thus retaining the separation of the map edges and preserving the asymmetry introduced by a rectangular map

## 4.3 Number of Clusters

Perhaps the most controversial part of cluster analysis involves determination of cluster number [47, 48, 49, 50]. One popular approach repeatedly samples single linkage hierarchical cluster trees generated by removing one or more data elements. Nodes that occur most frequently

in the sample trees define the number of clusters. Alternative methods use a cubic clustering criterion [51, 35] to estimate cluster number by minimizing the within cluster sum of squares while using standard statistical tests to determine the significance of error reduction. More recent methods evaluate statistical significance of cluster number by evaluating the distribution of correlations in a large number of randomized trials [23, 52]. The approach used here calculates the dependence of the S/F correlation on cluster size, and the uses the percent of maximal clustering to determine cluster number.<sup>2</sup> Using the ExMOA data set, SOM clusters were generated for a range of map dimensions, and the results are displayed in Figure 4. Based on this result, a cluster number above 110 is sufficient to achieve at least 99% coverage of S/F correlations. Our selection of 153 (17x9) clusters exceeds this criterion. *Fig. 4.*

#### 4.4 Number of Cell Lines

Our analysis explored the role of number of tumor cell lines in our SOM analysis using our measure of S/F correlation. The S/F correlation with varying the number of cell lines, shown in Figure 5, has two or three basic regimes. Below  $\sim 20$  cell lines the S/F correlation drops off dramatically, between 20-50 cell lines the correlation rapidly increases, while for greater than 70 cell lines the correlation achieves a maximum. Although further analysis of this result will not be presented here, there is a clear indication that a near optimal clustering result can be achieved with fewer than the 80 tumor cell lines analyzed herein. *Fig. 5.*

#### 4.5 Robustness

We have investigated the behavior of the SOM method towards noisy and degraded data. Figure 6, Panel A shows a sigmoidal decrease in S/F correlation with decreasing completeness of the input data. This data set was degraded by systematically removing data elements with the most extreme Z-score values. The results show that from 100% thru 70% completeness of data the S/F correlation is resistant to this degradation. Below 70% the correlation coefficient rapidly decreases; approaching a minimum at 0.45. This analysis illustrates the relation between *Fig. 6.*

---

<sup>2</sup>In SOM clustering cluster size is equivalent to map dimensions. Percent of maximal clustering measures closeness to the asymptotic limit.



strong signals, as measure by extreme values of absolute deviation Z-scores, and the quality of the clustering. The behavior of the S/F correlation with degraded data is relatively stable against data sets which exhibit greater than a 10% coefficient of variation (see Figure 6, Panel B). Below a Z-score of  $\sim 1.1$  absolute deviation units the S/F correlation is drastically decreased. Consistent with intuition, a data vector with a large amount of diversity (stronger signal) can be more easily assigned to a cluster, when compared to data vectors with a small absolute deviation. Based on this result, our analysis excludes data vectors with an absolute deviation below 8%.

## 5 COMPLETE MAP

The SOM analysis of the MOA and ExMOA data sets established important guidelines for analyzing all the publicly available data for synthetic compounds screened against the NCI's tumor cell panel. To our knowledge a simultaneous analysis of the complete screening data set has not appeared in the literature. As our analysis will demonstrate, the ability to simultaneously analyze this data offers a unique perspective into the complete range of biological response patterns for these tumor cells and provides valuable information for extracting additional details about cellular activities, assessing similarities and differences in response patterns for these activities and determinations of unique and under explored regions in cellular response space. The initial screening data set included measures of cytotoxicity for  $\sim 33K$  (32,918) compounds. Filtering of this data, based on retaining only those data vectors with greater than 8 percent absolute deviation, reduced this set to  $\sim 20K$  (19,867) compounds. The SOM analysis generated a map with dimensions 41 rows by 26 columns, to yield 1066 map clusters (see Figure 7).

*Fig. 7.*

The SOM map represents an alternative to the more classical use of dendrograms for displaying cluster results. As was mentioned earlier, the anti-tumor drug screening data set does not appear to lend itself to hierarchical organization, where each clade's position is assigned only on the basis of its closest neighbor. The two-dimensional SOM representation allows clusters to have upwards of 6 neighbors, for the hexagonal representations used here. The

distance between neighboring map loci<sup>3</sup> is indicated by the connecting color (in figure 7, close in distance is shown in red, far apart in violet.) As an alternative to presenting the details of each of the 1066 clusters on this map, fifty regions have been defined that group individual map nodes with the most similar response profiles. These regions are similar to individual clades of a hierarchical dendrogram, however their organization is not restricted to the simple division into groups. Using this convention, clusters on the complete DTP map can be assigned to six functional categories according to the apparent cellular activity of the compounds within each of these six functional clusters. Using a mnemonic convention, we identify six classes of cellular activities: mitosis(M), nucleic acid synthesis(S), membrane transport and integrity(N), phosphatase and kinase mediated cell cycle regulation(P) and two remaining regions we arbitrarily have labeled Q and R. Justifications for these broadly assigned cellular activities will be obtained solely from literature sources. Despite these rather general assignments of cellular activity, this convention serves to organize the large amounts of data in this screen. Experimental validations of these putative cellular activities, in cases where actual molecular targets have been identified, will be reported. However, in the absence of published reports, these classifications should be treated as speculations that serve as hypotheses for further experimental testing. While the absolute boundaries between classes, as far as map position and the division into six classes are somewhat arbitrary, they provide a useful framework for discussion of the anti-cancer drug screen results. It should be noted that the mapping, clustering, and other derived data are independent of the division into classes and regions of the map.

Following our mnemonic convention, Region M consists of two subregions,  $M_1$  and  $M_2$ , located at the top left corner of the map. Moving diagonally from this corner towards the lower right-hand corner, are Regions P (subregions  $P_1$ - $P_{14}$ ) and N (subregions  $N_1$ - $N_{12}$ ), above and below the diagonal, respectively. Region S (subregions  $S_1$ - $S_7$ ) appears at the lower right portion of the map, separated from Regions M, N and P by Regions Q (subregions  $Q_1$ - $Q_7$ ) and R (subregions  $R_1$ - $R_9$ ). Figure 8A shows the projection of 171 standard anticancer agents that have been clinically evaluated [53, 54, 55] consisting largely of antimetabolic compounds which are located in region M of the complete DTP map, and agents that affect nucleic acid

*Fig. 8*

<sup>3</sup>The 1066 map clusters will be referred to as loci, and will be identified by their row and column coordinates. This convention will provide a convenient reference to map positions.

biosynthesis which are, with a few exceptions, located in Region S. Below, we will more fully characterize Regions M and S and portions of Regions P and N for cellular activity by presenting a comprehensive analysis of within-cluster members and literature reports of their putative cellular actions. To facilitate reporting of these results, map loci containing compounds with similar cellular activity, will be identified by color-coded hexagons, placed at their respective map positions as shown in Figure 8B. The reader interested in all 1066 clusters on the complete DTP map may solicit our web site at <http://spheroid.ncifcrf.gov>.

## 5.1 Region S: Nucleic Acid Synthesis

Considerable research interest has been directed into selectively targeting tumor cells using antimetabolites of purine and pyrimidine nucleotide metabolism [56, 57, 58]. Three metabolic targets are found within the purine and pyrimidine class of antimetabolites: 1) agents that interfere with the synthesis of RNA and DNA precursors, 2) direct inhibitors of DNA synthesis and 3) compounds that are incorporated into RNA and DNA which later disrupt cellular processes. The first class of inhibitors includes the antifolates and inhibitors of rate-limiting pyrimidine and purine de novo synthesis enzymes. These compounds are located in subregions  $S_4$  and  $S_5$ , and appear as purple and red hexagons in Figure 8B.

**Antifolates:** More than fifty dihydrofolate reductase (DHFR) inhibitors are all locally clustered in subregion  $S_4$  at the loci k36.21, k37.21, k37.22 and k38.20<sup>4</sup>. This list of tetrahydrofolate (THF) analogs include methotrexate, trimetrexate, triazinate, methasquin, piritrexim isethionate, pyrimethamine, PT523, and pyrimethamine. These antifolates are believed to function by binding to DHFR, disrupting the thymidylate synthase(TS) complex which results in the depletion of dTMP pools and, thus, inhibiting de novo pyrimidine nucleotide biosynthesis. In addition, certain THFs act as competitive inhibitors by directly binding to folate-dependent enzymes, such as TS, and interfering with co-factor binding.

The complete DTP map further separates the THF antifols into two overlapping classes:

---

<sup>4</sup>Each locus in the SOM maps is identified by its coordinate locations, row.column. For example the cluster at locus row=10 and column=20 is referred to as k10.20. This convention will be adopted throughout the text.

those which use the reduced folate carrier pathway, and those which can transverse the cellular membrane (by potocytosis via folate binding proteins) and/or through passive diffusion [57]. The classical antifolates, such as methotrexate (MTX), contain a glutamic acid tail and therefore require the reduced folate carrier pathway. Mutations in this pathway are a common cause of antifolate resistance [57]. The class of tetrahydrofolates which contain the glutamic acid tail are located chiefly in subregion  $S_4$  at the loci k37.22 and k38.20. By contrast, THFs which do not contain the glutamic acid tail, such as the small lipophilic compound pyrimethamine, are located in subregion  $S_4$ , at k36.21 of the complete DTP map (see Table II.) These non-glutamate compounds retain their ability to kill cancer cells even in the presence of an inactive reduced folate carrier pathway. Therefore, cell lines that have mutations in the reduced folate carrier pathway are insensitive to MTXs while still sensitive to the pyrimethamines. The data vectors for the clusters k38.20 and k36.21 are consistent with these differences in cell sensitivity; their tumor cell patterns differ chiefly in the response of K-562 human leukemia cells, a tumor cell line that is known to be very sensitive to the pyrimethamines, while insensitive to MTXs: a result consistent with the above described differences in cellular transport and other related alterations such as target enzyme amplification and reduced cellular capacity for polyglutamation. Table II displays examples of compounds located at each of the regions in the complete DTP map.

*Tbl. II*

**UMP Biosynthesis:** An important class of anti-cancer agents inhibit biosynthesis of the DNA/RNA precursor uridine monophosphate(UMP). These agents exert their cytotoxic mechanism by depletion of UMP pools thereby halting DNA/RNA synthesis. Upstream of the biosynthetic synthesis pathway of UMP are the enzymes carbamoyl phosphate synthetase II, aspartate transcarbamoylase, dihydroorate dehydrogenase, OPRTase, and orotidyl decarboxylase [57]. Our complete DTP map locates inhibitors of these enzymes in subregion  $S_4$  chiefly in loci k35.22, k36.22, and k37.22, identified as red hexagons in the lower right portion of Figure 8B; as would be expected for DNA/RNA antimetabolites. Specifically the enzymes/inhibitors are: carbamoyl phosphate synthetase II/acivicin (at k35.23 and k38.20), aspartate transcarbamoylase/PALA (at k35.21), dihydroorate dehydrogenase/brequinar (at k36.22 and k37.22), dihydroorate dehydrogenase/dichloroallyl lawsone (at k35.22), and orotidyl

decarboxylase/pyrazofurin (at k36.22).

**CTP-S:** Cytidine triphosphate synthetase (CTP-S) is an important target for cancer chemotherapy. Compounds that affect pyrimidine biosynthesis by inhibiting cytidine triphosphate synthetase conversion of UTP to CTP are found centered at locus k38.20, at the bottom of subregion  $S_4$ , shown as red hexagons in Figure 8B. Some of the well characterized chemical agents which inhibit CTP-S are cyclopentylcytosine (at k37.19, k37.21, k38.19 and k38.20), 3-deazauridine (at k36.22 and k41.20), gemcitabine (at k39.25 and k41.24) and acivicin (at k35.23 and k38.20).

The CTP-S inhibitor, acivicin, also has activity against carbamoyl phosphate synthetase II, as mentioned above. This behavior is common among nucleic acid affecting agents. Many of the chemical compounds which affect DNA/RNA biosynthetic enzymes have multiple mechanisms of action owing to their nucleic acid moieties. We have observed the convergence in the clustering of nucleic acid agents with increasing test concentration. An example of this behavior can be seen in subregion  $S_4$  at loci k37.21, where high test concentrations of antifolates (MTX's and pyrimethamines), CTP-S inhibitors, and IMP dehydrogenase (IMPDH) inhibitors are all co-clustered. At lower concentrations, where the inhibitory effect results in more selective targeting, the various classes of chemical agents are separated into clusters which better reflect their different molecular targets.

In general, individual loci of the map appear to represent cellular response mediated predominantly by a single molecular target or pathway. Other loci apparently are comprised of compounds with cellular activity due to joint action at multiple targets. In certain situations experimental evidence is available to document concurrent operation of multiple mechanisms to create "overlapping phenotypes of multidrug resistance" [59]. Thus far, we have only examined a subset of the 1066 map loci. It is likely that there exist map loci which cannot yet be associated with any molecular mechanism which represent novel and unknown anti-cancer drug classes, it is also likely that loci exist which represent a combination of activities similar to overlapping phenotypes resulting from drug resistance.

**TS:** The clusters associated with compounds active against thymidylate synthase(TS) are

centered at subregion  $S_5$  in the neighborhood of k40.20, located as red hexagons in Figure 8B. The TS affecting clusters are dominated by 5-fluorouracil (5-FU) and structurally related compounds. 5-FU's cytotoxic mechanism is via TS inhibition and through direct incorporation into RNA and DNA. The thirteen other compounds found co-clustered with 5-FU are all fluorinated uracil and polyethylene glycol conjugates with molecular weights ranging from 1480 to 6640 Daltons. The location of this group of clusters, adjacent to the CTP-S inhibitors at k41.20, suggests that similarities in their cellular response patterns accompany defects in interconversion between uridine phosphate moieties (UTP and dUMP) to cytidine(CTP) and thymidine(dTMP) species, respectively.

**Purine Biosynthesis:** De novo purine biosynthesis is another target for anti-cancer agents. A number of clinically important drugs inhibit the formation of inosinate monophosphate, a precursor to the formation of purine nucleotides. These compounds are located in the lower right portion of Figure 8B as purple hexagons. Compounds which inhibit the IMP synthesis pathway enzyme, glutamine amidotransferase, are centered at locus k37.20 in subregions  $S_4$  and  $S_5$ . These agents include 6-diazo-5-oxo-L-norleucine (DON, at k37.21 and k38.19), L-azaserine (at k37.20), AT 125 (at k35.23 and k38.20), and hydroxy AT 125 (at k37.20). Inosinate monophosphate dehydrogenase (IMPDH) is an important enzyme in the de novo biosynthesis of purine nucleotides. Agents that are purported to inhibit IMPDH by binding to its NAD<sup>+</sup> co-factor site, tiazofurin and selenofurin [60], are found in Region  $Q_5$  at loci k38.13 and k37.14 (see Table III.) A related agent, cordycepin, is also found at k37.14, suggesting a different molecular target for tiazofurin and its analogs [61]. In direct contrast to the tiazofurin response, the NAD<sup>+</sup> analog, mycophenolic acid, is located in Region  $S_4$  at k37.20, along with the IMP upstream inhibitors described above; a result consistent with its role as a purine anti-metabolite. *Tbl. III*

In addition to the NAD<sup>+</sup> mimetics described above, a number of agents inhibit IMPDH by binding to its purine pocket. These purine analogs, such as 6-mercaptopurine and 6-thioguanine, are also located in Region  $S_6$  around the neighborhood of k34.17 and k35.18. As observed with the IMP upstream inhibitors, IMPDH and mercaptopurines, agents that affect purine biosynthesis or exogenous nucleic acid incorporation, are found on the complete

DTP map in subregions spanning  $S_5$ ,  $S_6$  and  $S_7$ ; as would be expected for nucleic acid anti-metabolites.

**Ribonuclease Reductase:** A functioning ribonuclease reductase(RR) is necessary for the biosynthesis of both deoxypurine and deoxypyrimidine nucleotides. To perform its role of reducing nucleotides, RR requires a functional iron center. The complete DTP map separates RR inhibitors into two classes: (i) Hydroxyurea and similar compounds pyrazoloimidazole (IMPY), and guanazole; located in subregion  $S_3$  at k39.25, identified by the red hexagon in subregion  $S_3$  of Figure 8B, from (ii), compounds that act by chelating metals; located at k32.22 in subregion  $S_2$ , shown as the single light red hexagon in Figure 8B. The metal chelators include deferoxamine, and terpyridine and its analogs (see Table IV.) It is interesting to note the co-clustering, *Tbl. IV* at k39.25, of hydroxyureas with the DNA polymerase alpha inhibitor aphidicolin and its derivatives. This location, near the lower right edge of the complete DTP map, is characteristic of DNA replication inhibitors such as anti-topoisomerases and alkylating agents; to be discussed further in the following sections. Inhibition of DNA polymerase has also been reported with deoxycytidine analogs such as cytarabine (ara-C). Cytarabine is found at locus k36.24, along with its derivatives adamantoyl cytarabine, palmitoyl cytarabine and fazarabine [62] and other similar compounds.

## 5.2 Subregion $S_3$ : Topoisomerase Inhibitors

DNA topoisomerases are enzymes which catalyze DNA strand breaking and unwinding during cellular replication and RNA transcription [63, 58]. In eukaryotic cells, DNA topoisomerase I (topo I) and topoisomerase II (topo II) each perform similar but distinctly different roles in DNA unwinding. Topo I catalyzes single strand 'nicking' which allows supercoiled DNA to unwind, and is relatively constantly expressed during the cell cycle [57]. In contrast, topo II is most highly expressed at the end of the S phase of cellular replication and throughout the G2 phase, to facilitate chromosomal duplication by relaxing and unwinding the DNA duplex [57]. The molecular mechanism of DNA unwinding also differs between these two enzymes; topo I nicks a single strand to allow DNA unwinding, whereas, topo II uses ATP to pass one DNA strand through its complementary strand and then rejoin the 'break'. SOM clustering of compounds

active as inhibitors of topo I and II are found in separate, but adjacent regions in the nucleic acid portion of the complete DTP map. Topo I inhibitors, consisting of the camptothecin analogs, are found in the lower part of subregion  $S_3$ , within adjacent loci at k40.24 and k41.24, identified as dark brown hexagons in Figure 8B. Topo II inhibitors are located in two overlapping groups: the 'etoposide' group clustered in subregion  $S_3$  at loci k38.26, k39.26, k40.24, k40.25, and k40.26; and, the 'anthracycline' group also in subregion  $S_3$  at loci k41.25 and k41.25, shown as light brown hexagons in Figure 8B.

Other agents similar to these topoisomerase inhibitors include the bleomycin family of DNA damaging antibiotics. These compounds are found in Region  $S_3$  at locus k40.22, appearing as a dark brown hexagon in Figure 8B. In addition, the classes of DNA intercalators, ellipticine and methoxyellipticine, imidazoacridiones (e. g. NSC645812), and triazoloacridiones (e. g. NSC645827) centered around k41.16 in subregion  $S_7$ , and the acridines, such as quinacrine mustard, near k41.17 and k41.18, are shown in Figure 8B as green hexagons as the bottom of the map.

**Bisantrene:** While bisantrene is regarded as a putative topoisomerase inhibitor, its cluster position is not associated with other topoisomerase inhibitors. Rather it is found in the mitotic/membrane portion of the complete DTP map, in subregion  $M_2$ , at k11.3. This result suggests that under the conditions of the DTP cell-line screen, bisantrene does not function as a topoisomerase inhibitor. Confidence in the reliability of bisantrene's cluster location is supported by its three screening measurements at differing test concentrations [ $\log(\text{highest test concentration}) = -5.0, -4.0$  and  $-3.6$ ]; all of which cluster at k11.3. A number of bisantrene analogs are also found at k11.3, which include the bis nitrogen chain containing anthracenes, anthracyclines, and acridines. Other well characterized antibiotics, co-clustered at k11.3, are puromycin, an actinomycin D derivative, and tubulosine. The actinomycin D derivative has recently been implicated as a blocker of Grb2-SH2 access to the Shc/Ras and Shc/phosphatidylinositol 3-kinase pathways (PI3K) [64]. Additional support for this role is found in neighboring clusters, which include other anti-neoplastic antibiotics, such chromomycin A3. The following sections will amplify the putative action of anti-neoplastic antibiotics on the NCI's panel of tumor cells.



**Fostriecin:** The compound fostriecin has also been identified in the literature as a topo II inhibitor [65] based, in part, on similarities in tumor cell responses when compared to other nucleic acid affecting agents [54, 57]. The SOM analysis, however, places fostriecin in the portion of the complete DTP map associated with kinase and phosphatase mediated cellular regulation (subregion  $P_{13}$ ) at cluster k6.8. Recent literature confirms this placement with reports of fostriecin activity against protein phosphatase 2A [66, 67], while simultaneously demonstrating a lack of topo II activity [68]. Additional inspection of agents in k6.8 further supports this activity by co-clustering the compound cytosstatin, which is known to inhibit cell adhesion to the extracellular matrix by selectively inhibiting protein phosphatase 2A [69]. This observation illustrates the power SOM clustering and its application to the complete data set available in the NCI's tumor screen. The ability to inspect the global response of tumor cells to these agents provides information about locally clustered compounds, which, in turn, can be used to assess cellular activity by comparisons to inhibitors of known molecular targets.

### 5.3 Subregion $S_3$ : Alkylating Agents

Alkylating agents introduce exogenous covalent bonds in nucleic acids and associated proteins which later interfere with transcription and translation. These agents are clustered adjacent to the topo II inhibitors, consistent with their similar mode of action, and appear as orange-brown hexagons at the lower right of Figure 8B. As observed for inhibitors of nucleic acid biosynthesis, cluster separations are also observed within families of alkylating agents, according to their functional subtypes. Alkylating agents with a bi-functional electrophilic leaving group, such as busulfan, thio-tepa, chlorambucil, and di-platinum compounds, are all found in subregions  $S_1$ ,  $S_2$  and  $S_3$ . The nitrosourea alkylating agents, asaley and the mitomycins (which are antibiotic alkylating agents) are found in subregion  $S_3$  at k38.23 and k41.25. Bi-functional alkylating compounds that contain two platinum atoms are found in neighboring clusters at k37.25 and k38.25. The monoplatinum compounds are found in four main clusters which differ in their chelating moieties: diaminocyclohexyl at k39.19, diaminoamino at k39.23, bis(aminomethyl)cycloalkylsilyl at k35.25, and imidazolyl at k32.26. A clear exception is found with the class of nitrosourea alkylating agents; CCNU, methyl-CCNU, BCNU, and

cis-4-Hydroxy CCNU. These compounds are not found in Region S, but in subregions  $N_5$  and  $N_2$  at k22.6 and k23.6, respectively. This result suggests a cellular activity different from DNA alkylation, possibly by alkylating proteins within regulatory cellular pathways. The compounds that co-cluster with these CCNU's are phenazinomycin and 2-hydroxy-4,6-dimethylchalcone, the latter compound derived from the parent chalcone molecule, which, itself, is clustered with the DNA-methylation agents in the nucleic acid portion (Region S) of the map. This represents an example where the mode of action of the parent compound has been substantially altered via chemical modification.

## 6 Regions M and N: Mitosis and Cellular Membranes

Agents that impact arrest in the mitotic phase of cell division are found in the upper left portion of the complete DTP map in Region M, identified in Figure 8B by three differently shaded blue hexagons. The bulk of compounds in Region M interfere with the microtubule/actin cellular framework, consisting of taxanes (at k9.1 and k9.2), colchicines (k6.1), vinca alkyloids (k7.1), trimethoxystilbenes and peltatins (at k1-3.1-5) and other compounds such as the macrolide rhizoxin (k5.1) and nocodazole(k1.1) (see table V.).

*Tbl. V*

Compounds that affect cellular membranes are found in the Region N. In particular, at the left edge of subregion  $N_8$  and the bottom of subregion  $M_2$  are cationic surfactants that appear to directly act on the lipid bilayer to disrupt cellular membranes (shown in the light green hexagons at the upper left edge of Figure 8B.) Typically these compounds contain a positive charge and a strong hydrophobic moiety. An example would be a positively charged nitrogen embedded in fused planar aromatic rings. Additional examples of cationic surfactants are located in nodes along a line extending from k13.1 through k18.1 in Regions  $M_2$ ,  $N_8$  and  $N_4$ . These clusters include the arylphosphoniums, arylquinoliniziniums, dequaliniums, berberins, and charged ellipticines (e. g. NSC39310, NSC166454, NSC5355, and NSC264137 respectively.) Separation of tumor responses on the basis of charged and uncharged ellipticines has been previously reported by Shi et. al [33]. Other compounds in this region include a series of dihydroxyanthracenones at cluster k15.2, also known to have antipsoriatic activity via 5-lipoxygenase or other biosynthetic enzymes [70]. However, the location of these compounds

in the N region suggests a role in damaging membranes via generation of oxygen-radicals; a well known side-effect of some antipsoriatic agents. This is an example where the cell screen map can be used for drug discovery; via, for example, locating anthralin-like compounds not found in the membrane damaging region. Validation of this possibility could lead to the discovery of new antipsoriatic agents with reduced side effects.

Another mechanistic class of membrane targeting agents are ion channel inhibitors, located in Region  $N_{10}$  of the complete DTP map, shown as yellow hexagons in Figure 8B. Examples include the  $K/Ca^{2+}$  channel blockers, pimozide, at k9.7, verapamil, at k9.8 and k11.8; the tetrandrine class of calcium channel blockers, at k8.7, k7.8, k6.6; including tetrandrine, fangchinoline, oxyacanthine, funiferine, dauricine, at k11.6, chloroquin diphosphate, at k9.7 and prazosin, at k7.4, all located in subregion  $N_{10}$  (see Table VI.) Antihistamine H1 antagonists, like diphenylhydramine, are also found in this region, at k11.8. Their cellular activity is likely to perturb ion levels, while acting as H1 agonists [71], or as ion channel inhibitors [72, 73]. Another class of compounds mapped to subregion  $N_{10}$  is the phenothiazine family of dopamine antagonists (prochlorperazine, fluorophenothiazine, trifluopromazine, and clopenthixol) and the compounds pimozide (k9.7), metoclopramide (k9.12), and the spiperones (k8.11). Mapping these compounds to subregion  $N_{10}$  is most likely due to their effect on ion homeostasis. Tbl. VI.

It is instructive to note that the thioxanthene, lucanthone, which is generally accepted to act as a topo II inhibitor, also maps to subregion  $N_{10}$  at k10.7. Its location in Region N, and its structural similarity to phenothiazines, suggests action as a membrane disrupting agent under the conditions of the DTP cell screen (see Table VI.) Moreover, thioxanthenes which contain a hydroxyl moiety, such as hycanthone, are projected in Region S. These thioxanthenes are further subdivided into 7-hydroxy compounds in subregion  $S_3$  at k39.26 and 4-methoxy compounds  $\sim$  k39.16 in subregion  $S_7$ . Subregion  $S_3$  corresponds to topo II inhibitors, such as menogaril and hydroxydaunorubicin, while subregion  $S_7$  includes compounds belonging to the planar aromatic class of intercalators.

Agents known to affect the Golgi complex are mapped to Region N and are displayed in Table VII. Examples include brefeldin A and its analogs; found in subregion  $N_{11}$  at k12.10. Other compounds known to disrupt the Golgi complex are found in loci roughly along a line extending from k11.12 through k28.2 through the middle of Region N, shown as maroon hexagons Tbl. VII.

in Figure 8B. These agents include cytochalasins D,E,H (k11.12), okadaic acid (k16.6 and k10.1), limaquinone (k17.8), ilimaquinone (k19.5), avarol (k21.5) and cytochalasin A (k24.2 and k28.2). In addition to okadaic acid found at k10.1, the rough endoplasmic reticulum agent, thapsigargin, which targets  $\text{Ca}^{2+}$  transport molecules, is found at k9.3 [74, 75]. We note the correspondence between agents that affect the Golgi complex and those that disrupt the actin cytoskeleton. Co-clustered with cytochalasins D,E,H at k11.12 are the cucurbitacins, dolastatin 11, a jasplakinolide, and pectenotoxins, which are known to disrupt the actin cytoskeleton [76, 77, 78, 79]. The mechanistic coupling between the Golgi complex (and other membrane organelles), actin, and ion channels has only recently been indentified [80, 81, 82]. These published findings are consistent with their SOM map location in subregion N10, which is at the confluence of the mitotic, membrane and cell-cycle regions (see Figures 7 and 8B.)

## 7 Region P: Cellular Regulation and Apoptosis

Kinases and phosphatases associated with apoptosis and cell cycle regulation are mapped to the upper portion of the complete DTP map in Region P (see Figure 7.) Selected compounds in this region are also listed in Table XIII. Within this large portion of the map are compounds that function as protein kinase C activators, in subregion  $P_{13}$ , at k6.9, shown as pink hexagons near the top of Figure 8B. These compounds include the phorbols (mezerein, huratoxin, and prostratin) [83], gnidimacrin [84], and cytoblastin [85]. Adjacent to k6.8 is the previously described class of protein phosphatase 2A inhibitors, fostriecin and cytostatin [66, 69]. As noted earlier, these compounds have often been considered as topo II inhibitors [86, 54], however, literature now supports their role as a protein phosphatase 2A inhibitor. The neighboring cluster, k7.11, contains miltefosin and similar long chain alkylphosphocholines, such as perifosine. The location of this cluster, away from sub-regions  $N_8$  and  $M_2$ , suggests a companion activity in cell cycle regulation; consistent with literature reports supporting a cell cycle role for alkylphosphocholines [87, 88]. KRN5500 [at  $\log(\text{high test concentration}) = -8.0$  M] is clustered with the alkylphosphocholines in subregion  $P_{12}$  at k7.11. KRN5500 and the alkylphosphocholines share a long lipid chain and have potentially positively charged 'head' groups. KRN5500 has been shown experimentally to affect the Golgi complex [89]. It is likely that KRN5500 shares

*Tbl. XIII.*

a dual mechanism of action that includes the alkylphosphocholine's molecular target and the Golgi complex. Two of the three data vectors for KRN5500 are clustered in the Golgi disrupting region of the map, near brefeldin A, in subregion  $N_{11}$  at k11.14 [at  $\log(\text{high test concentration}) = -7.0$  and  $4.3$ ]; also co-clustered with its structural analog, septacidin.

Other kinase active compounds in subregions  $P_{12}$  and  $P_{13}$  include: bryostatin 1, at k4.11, which targets protein kinase C[90, 91], forskolin, at k5.12, which induces cAMP-like kinases[92], wortmannin at k5.12, a phosphoinositide 3-kinase (PI3K) inhibitor[93], and ursolic acid, at k4.11, which is involved in caspase activation and down regulation of the apoptotic protective c-IAPs proteins[94]. It is interesting to note the co-clustering of forskolin and wortmannin at k5.12 and the near-by clustering of the phorbol esters (k6.8). Forskolin is known to up-regulate cAMP and enhance kinase activity and exert its influence via ion channels [95, 96] while Ecay et. al. [97] reports on the wortmannin inhibition of forskolin-stimulated chloride secretion. Similarly, the results of Yamashita et al. [98] show that forskolin and phorbol esters to have opposite effects on the expression of mucin-associated sialyl-Lewis(a) in pancreatic cancer cells. Thus, exploration of co-clustered compounds may further illuminate known mechanistic pathways involved in cancer chemotherapy and reveal previously unknown molecular interactions.

## 7.1 Subregions $P_1$ And $P_2$ : Cyclin Dependent Kinase Inhibitors

Subregions  $P_1$  and  $P_2$  are identified as rich in cyclin dependent kinase (CDK) inhibitors. Included in these regions are the staurosporines (including UCN-01), quinazolines, paullones, flavopiridols, quercetins, and others. Localization of these compounds to this map region can be attributed, partly, to a high sensitivity on these compounds to the small cell lung(SCL), central nervous system(CNS), and renal(REN) cell panels, while at the same time exhibiting an insensitivity to the leukemic(LEU) and colon(COL) cell lines. Our analysis finds that Region P has been less completely probed when compared to Regions M and S [99]. The most extensively explored compounds in this Region P include  $\sim 100$  flavopiridol/quercetin analogs,  $\sim 15$  staurosporine/UCN-01 analogs, and  $\sim 10$  paullone analogs. In contrast, other compounds in this region, that may also function as CDK inhibitors, are represented by just a few measurements. Further generalization about these sparsely sampled regions in response space, beyond

our suggestion that they may potentially target cyclin dependent kinases, must await additional experimental verification.

A contributing factor in assessing the cellular activity of poorly investigated spaces in Region P is the possibility of multiple cellular targets. This is often true for small molecules, such as the flavopiridols, and especially true for purine mimetics which inhibit CDKs by binding their ATP pocket, as with the quinazolones and olomoucine analogs described above. In these situations the precise identification of cellular activity is difficult, however, the co-clustering of families of compounds provides evidence of their cellular activity. For example, the relatively large number of compounds with CDK inhibition activity found in subregion  $P_2$  supports its classification as a CDK inhibitor region(see Table IX.). Such is the case for families of staurosporines, rapamycins, and celphalostatins. The antineoplastic antibiotic rapamycin and its analogs are found in subregion  $P_2$  of the complete DTP map, at k11.26. Rapamycin's role as a cyclin dependent kinase modulator, based solely on its map location, distinguishes it from other antibiotics that function to inhibit nucleic acid synthesis, mitosis or membrane function. Rapamycin has recently been implicated as an inhibitor of mTOR kinase (also called FRAP and RAFT1), which is an important regulator of cell cycle progression and growth, as part of the PI3K pathway [100, 101]. The large number of co-clustered rapamycin analogs exhibit sufficient structural similarities to suggest a common cellular target, i.e. mTOR. Another notable cluster in the  $P_2$  subregion includes the cephalostatins 1-9 at k14.26. This unique co-clustering of these disteroidal alkaloid analogs supports our inference that their cellular activity results from targeting a cyclin-dependent kinase. Recent work suggests that cephalostatins can function as CDK4 kinase inhibitors, with a moderate activity for cephalostatin 1 of 20  $\mu$ M [102]. Co-clustering of structurally similar compounds appears to substantially strengthen this hypothesis about cellular targets. Such examples may motivate additional investigations in the more sparsely populated portions of Region P. Additional classes of steroidal compounds are found throughout the P, N, and M Regions of the complete DTP map. Table X lists a number of steroidal compounds found in Regions P and N, as well as esterdiols, found in Region M, at k1.4. Recent literature now supports the cellular activity of estradiols as anti-tubulin agents [103], consistent with its placement in Region M of our map.

## 7.2 Subregion $M_2$ : Antineoplastic Antibiotics

A large variety of antibiotics are known to have anti-tumor activity. Approximately a quarter of these antibiotics are placed in Region S of the complete DTP map, while the remaining compounds are in subregion  $M_2$ . Antibiotics located in Region S appear to act as nucleic acid and protein synthesis inhibitors to yield cellular response patterns similar to the broadly defined classes of antimetabolites and alkylators. These similarities would be expected for antibiotics such as mitomycin, which themselves contain alkylating moieties. The larger fraction of antibiotics located in subregion  $M_2$  include the aureolic acids, harringtonines, quassinoids, terfenadines, trichothecenes, cyclosporines, bouvardins, didemnins, valinomycins and others. These compounds are listed in Table XI and appear at the left edge of Figure 8B as orange hexagons.. *Tbl. XI*

Their location at the intersection of subregions  $M_2$ ,  $N_9$  and  $N_{10}$  suggests their possible role as membrane disrupting agents and/or cell cycle kinase agonists and antagonists. Antibiotics that effect membrane ion transport include valinomycin, which acts directly on membranes by shuttling chelated ions [104] and compounds which indirectly disrupt membrane integrity such as the quassinoid, glaucarubolone, which inhibits plasma membrane-associated NADH oxidase [105] and thapsigargin, which inhibits an endoplasmic reticulum  $\text{Ca}^{2+}$ -ATPase [106]. The remaining antineoplastic antibiotics include agents identified in experimental studies to affect apoptotic and cell cycle pathways. An as example the trichothec mycotoxins have recently been identified to induce apoptosis by triggering the ribotoxic stress response pathway which activates kinases JNK/p38 to induce apoptosis [107]. Tetrocarcin A has been shown to directly inhibit the anti-apoptotic function of Bcl-2 [108] while actinomycin D binds the Grb2-SH2 complex [64]. Anisomycin also appears to activate stress pathways via activation of JNK and p38 [109]. Mapping these antibiotics to locations that include and border regions designated to have functionalities as membrane disrupting agents, ion transport affecting agents and cell cycle agents affecting apoptosis, raises interesting questions about relationships between these mechanisms and their roles in cell death. Previously anti-tumor antibiotics such as actinomycin D and mythramycin were thought act primarily by binding to DNA and inhibiting RNA and protein synthesis [57]. The location of these compounds in  $M_2$  suggests alternate hypothesis for their mode of action when compared to agents projected to the nucleic acid



synthesis affecting region  $S_3$ , such as streptonigrin and albacarcin V. Experimental validations of these possible cellular activities are the subject of future data-mining efforts.

## 8 Cellular Chemo-Sensitivities

Our SOM analysis provides quantitative details useful for relating these results to the goals of discovery, decision-making and explanation. The current screen is designed to identify relationships between chemical and cellular response space; the subject of the above material. Another class of questions, frequently the topic of previous analyses of the NCI's tumor screening data, involves coherence of response within and between tumor cell panels. Answers to this question are needed for the development of clinical strategies based on differentially expressed molecular targets within classes of tumors [110, 111, 112, 2, 113]. Interest in this area is also motivated by efforts to redesign the tumor cell screen towards either fewer numbers of tumor cells or to more tailored screens focused on sensitivity to specific chemotypes. Questions about cellular sensitivities can be addressed by simply inverting the design strategy presented in our current analysis, i.e. rather than deriving information about chemical similarities from cellular patterns, cellular similarities are now derived from chemical response patterns<sup>5</sup>. Previously, this question had been addressed using different subsets of the complete DTP database, and a variety of alternative clustering methods [39, 40, 33, 31, 35, 10]. The following section summarizes our efforts to better characterize the NCI's set of tumor cells that lend themselves to decisions of this type.

Response similarities within each tumor panel are directly assessed by correlations of their within-panel responses. Tumor panels exhibiting the highest correlations possess the most similar response patterns. Figure 9 displays the panel averaged pairwise correlation coefficients along its diagonal (Panel A). Our results find the LEU panel with the most similar response patterns, followed by the CNS and COL panels; whereas, the most variable response patterns are found within the OVA, BRE and LNS panels. These results are consistent with earlier analyses that separately co-clustered the LEU, CNS and COL panels, while substantially integrating the

*Fig. 9.*

---

<sup>5</sup>More technically, response space for  $N$  compounds across 80 tumor cell types is inverted to treat 80 cells across  $N$  compounds.



remaining seven tumor cell panels throughout different clusters [10, 35]. Measures of these between panel correlations are displayed as off-diagonal elements in Panel A. These results revealed modest positive correlations between the COL:LEU:SCL panels and the LNS:CNS panels, and weak negative correlations between CNS and the LEU and COL panels. The distribution means for the within and between panel correlations (cf. Fig 9 Panel B) are 0.22 and -0.02, respectively; neither significantly large enough to reject the hypothesis of zero correlation between these tumor cell panels. The notion that some tumor panels demonstrate coherent within-panel responses and that weak correlations could also be found between tumor panels suggested a reanalysis of the complete DTP data set according to tumor cell type. Figure 10 displays as hexagons each tumor panel's average (across 20K compounds) cellular response at all 1066 nodes on the complete DTP map, colored according to sensitive (red) and insensitive (blue) cellular activity; sized proportional to the magnitude of this activity. Consistent with our earlier observations, the LEU panel exhibits the most correlated within-panel responses, as reflected by the largely equally sensitive response at most node positions. Furthermore, concordance between patterns of blue and red regions in these images reflect the above noted positive correlations between the COL:LEU:SCL panels and negative correlations between the LEU and COL panels.

*Fig. 10.*

These results provide an explanation for previously observed clustering results among these tumor cell panels. Using sets of standard anticancer agents comprised of fewer than 200 compounds, Keskin [35] and vanOsdol [39] found reasonable co-clustering within selected tumor cell panels: results that generally concur with the most correlated tumor panels discussed above. More recently, an analysis based on ~1400 anticancer compounds found cluster memberships for these tumor cells to be quite different from previous analyses [10]. Inspection of Figure 10 offers an explanation for these differences. As noted earlier, the set of standard anticancer agents associated with inhibition of nucleic acid biosynthesis are located at the lower right portion of our map, in Region S. This region finds the LEU, LNS, REN and PRO panels to exhibit qualitatively similar chemical sensitivities to agents that affect this pathway. Biochemical reasons for these chemical sensitivities can be found in additional studies. As an example, cluster k40.25 contains topo II agents as well as agents active as DNA polymerase and ribonuclease reductase inhibitors. The tumor cell panels most sensitive to these agents are the renal

(REN) and non-small lung (LNS) panels. As discussed above, DNA polymerases are inhibited by ara-C and other deoxycytidine analogs. Analogs of cytosine, especially those modified at the 5-position, are known to reduce methylation of cytosine incorporated into DNA and lead to maturation of the lung tumor cell line A549 [114], a result consistent with the high sensitivity of the LNS panel to deoxycytidine analogs. In contrast, k37.22 consists largely of antifolate compounds. The most sensitive tumor panels in this cluster are the colon (COL) and the renal (REN) tumor cells. Antifolates are known to inhibit the purine biosynthesis folate-dependent enzyme glycinimide ribonuclease transformylase [58] and induce differentiation of the colon cell line HL-60 [115]. This effect may contribute to the enhanced sensitivity of the colon (COL) and renal (REN) cell lines to antifolates. The clear message of these examples, as well as the salient features of the ten tumor panels displayed in Figure 10, reveal the high dependence of cellular response on the selected chemotype. These responses are quite variable, depending on map location and tumor cell type, and are consistent with the notion that highly heterogeneous chemical probes elicit a wide range of responses within different tumor panels. Overall, however, these similarities do not appear to be strongly correlated when analyzed over the complete set of chemotypes tested in the screen. Numerous individual cases exist, however, where similar cellular sensitivities (and insensitivities) are apparent, depending on chemotype. For example, inspection of k1.11 identifies chemotypes with high cellular sensitivity to only the BRE panel, while k10.12 selectively identifies chemotypes that are insensitive only to the BRE panel. Once again, evidence that cellular response patterns are highly dependent on the chemotypes used to probe their sensitivity. A more detailed analysis of the nature of these similarities often finds that panel identity does not translate into response similarity. Explorations of the nature of these differences will be the subject of future analyses. In general, however, this finding indicates a limited capacity to associate chemotype with tumor panel response in these cell-based screens.

## 9 SUMMARY

A suite of internet accessible (<http://spheroid.ncicrf.gov>) computational tools has been assembled for analysis and visualization of large multivariate data sets. A range of important

analytical questions related to data analysis were addressed that found significant improvements in cluster quality could be realized from data conditioning procedures of Z-score normalization, capping and exclusion of missing data. Further analyses investigated the importance of completeness of cell lines in the screen and data degradation on cluster quality. These studies provide a foundation for clustering analysis of the complete set of publicly available tumor screening data. This analysis identified relationships between chemotypes of screened agents and their effect on four classes of cellular activities: mitosis(M), nucleic acid synthesis(S), membrane transport and integrity(N) and phosphatase and kinase mediated cell cycle regulation(P). Validations of these cellular activities, obtained primarily from literature sources, found i) strong evidence supporting within cluster memberships and shared cellular activity, ii) indications of compound selectivity between various types of cellular activity and iii) strengths and weaknesses of the NCI's tumor screen data for assigning compounds to these classes of cellular activity. Subsequent analyses of averaged responses within these tumor panel types finds a strong dependence on chemotype for coherence among cellular response patterns.

**Acknowledgment.** This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. NO1-CO-56000. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organization imply endorsement by the U. S. Government.

**Supporting Information Available:** Expanded version of structure tables II-XI including NCI identification numbers (also called NSC numbers), and a primer on understanding self-organizing map figures. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## 10 Appendix: Evaluation of Clustering Quality

Drug screening patterns produce a vector of growth inhibition for the panel of cells in the screen. These measures of diversity can range from no apparent differences to highly textured response patterns reflecting little, intermediate and high levels of growth inhibition within the panel of tumor cells. A considerable challenge to researchers analyzing this data is the problem of evaluating the quality of clustering results based on these measures of diversity. At the heart of this problem is the difficulty of evaluating whether measurement noise contributes significantly towards the occurrence of random statistical correlations [19, 36]. Oftentimes duplicate tests are run under the same experimental conditions and averaged results are presented [28]. This approach contributes to the reduction of measurement noise but does not rigorously quantify the noise components resulting from individual tests conducted at different times. Fortunately, large scale screening efforts, such as the NCI's tumor screen, often contain essentially duplicate data, in this case measuring compounds at slightly different test concentrations. These data sets provide a means to assess the quality of different clustering methods. A key underlying assumption in this analysis is that anticancer compounds measured at two slightly different concentrations should yield a similar growth inhibition pattern; an assumption that should remain true if each experiment is reproducible and dose dependent effects do not alter the underlying biological response. Also, this analysis excludes results from extreme cases where test concentrations are insufficient to either evoke a biological response or are conducted at test concentrations that are lethal for all cell lines; each case yielding no diversity in the biological response across cell types. Based on these assumptions, perfect clustering should locate compounds tested in replicate at similar concentrations in the same cluster. The analysis presented below will evaluate the performance of SOM clustering versus pairwise correlation analysis to place replicate concentration tests in the same cluster.

The measurements for the set of ExMOA compounds presented earlier were used in our analysis. Only data sets above the signal cut-off of 0.08 absolute deviation units across cell lines were selected. This set yielded 214 replicate concentration tests. To construct a randomized set of replicate concentration tests, all possible random pairings of compounds with replicate concentration data were generated ( $N=21,156$  random pairs). Random pairings were not

permitted between identical compounds or compounds which shared the same mechanism of action class. Table XII and Figure 11 displays the ability of the 3d MIND method and a standard pairwise method (COMPARE<sup>6</sup>) to separate the replicate concentration pairs from the randomized data set. Perfect clustering was found with the SOM method for 65 of the replicate pairs, located 5.0 standard deviation units away from the mean of the randomized data set. The leftmost open bar on Panel A in Figure 11 represents these pairs. An additional group of 10 compounds are found as nearest neighbors on the map at 2.5 standard deviation units from the randomized set. Within this set of compounds only 24 of the 21,156 random pairs were co-clustered with the replicate concentration pairs, to yield a false positive detection of only 0.11%. In total, the SOM method correctly identified 75 of the replicate concentration pairs while rejecting 98.89% of the randomized pairs.

*Tbl. XII  
Fig. 11.*

Clustering results based on pairwise correlation coefficients derived with the COMPARE program are shown in Figure 11 Panel B. In this case, overlapping Gaussian-like distributions characterized the results for the replicate concentration pairs and the randomized pairs. Above a correlation coefficient of 0.888, or at least 3.0 standard deviation units from the population mean of 0.31, 25 replicate concentration pairs were correctly identified, with only one false positive in this set. At 2.6 standard deviation units (correlation coefficient  $\geq 0.754$ ), 75 replicate concentration pairs were correctly found along with 279 of the randomized pairs, to yield a 98.58% rejection rate for false positives. This apparently small difference in rejection rate, however, translates into a large number of false positives; based on the set  $\sim 3 \times 10^4$  compounds in the screen a 1.31% relative difference in rejection rate between the two methods translates into  $\sim 400$  additional false positives for comparisons based only on pairwise correlation coefficients.

---

<sup>6</sup>The pairwise correlation analysis was completed with the COMPARE tool at <http://dtp.nci.nih.gov/docs/compare/cmpmatrix.html>. The analysis of random pairs was based on 11,210 unbiased samples selected from the total set of random pairs.

## References

- [1] Bellenson, J. Integrating information technology and drug discovery processes. *Nature Biotechnology* **1998**, *16*, 597–598.
- [2] Eisen, M. B.; Brown, P. T.; Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Nat. Acad. Sci.* **1998**, *95*, 14863–14868.
- [3] Weinstein, J. N.; Myers, T. G.; O'Connor, P. M.; Friend, S. H.; Fornace, A. J.; Kohn, K. W.; Fojo, T.; Bates, S. E.; Rubenstein, L. V.; Anderson, N. L.; Buolamwini, J. K.; vanOsdel, W. W.; Monks, A.; Scudiero, D. A.; Sausville, E. A.; Zaharevitz, D. W.; Bunow, B.; Viswanadhan, V. N.; Johnson, G. S.; Wittes, R. E.; Paull, K. D. An information-intensive approach to the molecular pharmacology of cancer. *Science* **1997**, *275*, 343–349.
- [4] Weinstein, J. N.; Kohn, K. W.; Grever, M. R.; Viswanadhan, V. N.; Rubinstein, L. V.; Monks, A. P.; Scudiero, D. A.; Welch, L.; Koutsoukos, A. D.; Chiausa, A. J.; Paull, K. D. Neural computing in cancer drug development: predicting mechanism of action. *Science* **1992**, *258*, 447–451.
- [5] Cox, B.; Denyer, J. C.; Binnie, A.; Donnelly, M. C.; Evans, B.; Green, D. V.; Lewis, J. A.; Mander, T. H.; Merritt, A. T.; Valler, M. J.; Watson, S. P. Application of high-throughput screening techniques to drug discovery. *Prog Med Chem* **2000**, *37*, 83–133.
- [6] Sundberg, S. A. High-throughput and ultra-high-throughput screening: solution- and cell-based approaches. *Curr Opin Biotechnol.* **2000**, *11*, 47–53.
- [7] Floyd, C. D.; Leblanc, C.; Whittaker, M. Combinatorial chemistry as a tool for drug discovery. *Prog Med Chem.* **1999**, *36*, 91–168.
- [8] Olsen, M.; Iverson, B.; Georgiou, G. High-throughput screening of enzyme libraries. *Curr Opin in Biotech* **2000**, *11*, 331–337.

- [9] Ostergaard, M.; Thykjaer, T.; Gromov, P.; Yu, J.; Palsdottir, H.; Magnusson, N.; Orntoft, T. F. Gene expression profiling: monitoring transcription and translation products using DNA microarrays and proteomics. *FEBS Lett* **2000**, *480*, 2–16.
- [10] Scherf, U.; Ross, D. T.; Waltham, M.; Smith, L. H.; Lee, J. K.; Tanabe, L.; Kohn, K. W.; Reinhold, W. C.; Myers, T. G.; Andrews, D. T.; Scudiero, D. A.; Eisen, M. B.; Sausville, E. A.; Pommier, Y.; Botstein, D.; Brown, P. O.; Weinstein, J. N. A gene expression database for the molecular pharmacology of cancer. *Nature Genet.* **2000**, *24*, 236–244.
- [11] Andrade, M. A.; Bork, P. Automated extraction of information in molecular biology. *FEBS Lett.* **2000**, *476*, 51–5.
- [12] Bassett, D. E. J.; Eisen, M. B.; Boguski, M. S. Gene expression informatics—it's all in your mine. *Nat. Genetics* **1999**, *21 1 Suppl*, 51–55.
- [13] Duggan, D. J.; Bittner, M.; Chen, Y.; Metzler, P.; Trent, J. M. Expression profiling using cDNA microarrays. *Nat. Genetics* **1999**, *21 1-Suppl*, 10–14.
- [14] Brazma, A.; Vilo, J. Gene expression data analysis. *FEBS Lett* **2000**, *480*, 17–241.
- [15] Bittner, M.; Meltzer, P.; Chen, Y.; Jiang, Y.; Seftor, E.; Hendrix, M.; Radmacher, M.; Simon, R.; Yakhini, Z.; Ben-Dor, A.; Sampas, N.; Dougherty, E.; Wang, E.; Marincola, F.; Gooden, C.; Lueders, J.; Glatfelter, A.; Pollock, P.; Carpten, J.; Gillanders, E.; Leja, D.; Dietrich, K.; Beaudry, C.; Berens, M.; Alberts, D.; Sondak, V. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **2000**, *406*, 536–540.
- [16] Staudt, L. M.; Brown, P. O. Genomic views of the immune system. *Annu Rev Immunol.* **2000**, *18*, 829–859.
- [17] Tamayo, P.; Slonim, D.; Mesirov, J.; Zhu, Q.; Kitareewan, S.; Dmitrovsky, E.; Lander, E. S.; Golub, T. R. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Nat. Acad. Sci.* **1999**, *96*, 2907–2912.

- [18] Meyer, E. F.; Swanson, S. M.; Williams, J. A. Molecular modeling and drug design. *Pharmacology and Therapeutics* **2000**, *85*, 113–121.
- [19] Sneath, P. H. A.; Sokal, R. R. *Numerical Taxonomy*. W. H. Freeman and Company: San Francisco, 1973.
- [20] Becker, R. A.; Chambers, J. M.; Wilks, A. S. *A language and system for data analysis*. Bell Laboratories Computer Information Services: Murray Hill, NJ, 1981.
- [21] Tavazoie, S.; Hughes, J. D.; Campbell, M. J.; Cho, R. J.; Church, G. M. Systematic determination of genetic network architecture. *Nat Genet.* **1999**, *22*, 281–285.
- [22] Meyer, R. D.; Cook, D. Visualization of data. *Curr Opin Biotechnol* **2000**, *11*, 89–96.
- [23] Alon, U.; Barkai, N.; Notterman, D. A.; Gish, K.; Ybarra, S.; Mack, D.; Levine, A. J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Nat. Acad. Sci.* **1998**, *96*, 6745–6750.
- [24] Brown, M. P. S.; Grundy, W. N.; Lin, D.; Cristianini, N.; Sugnet, C. W.; Furey, T. S.; Ares, M. J.; Haussler, D. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Nat. Acad. Sci. (USA)* **2000**, *97*, 262–267.
- [25] Kohonen, T. *Self-Organizing Maps*. Springer Verlag: Germany, 1995.
- [26] Toronen, P.; Kolehmainen, M.; Wong, G.; Castren, E. Analysis of Gene expression data using self-organizing maps. *FEBS Letts.* **1999**, *451*, 142–146.
- [27] Paull, K. D.; Shoemaker, R. H.; Hodes, L.; Monks, A.; Scudiero, D. A.; Rubenstein, L.; Plowman, J.; Boyd, M. R. Display and analysis of patterns of differential activity of drugs against human tumor cell lines: development of mean graph and COMPARE algorithm. *J. Natl. Cancer Inst.* **1989**, *81*, 1088–1092.
- [28] Monks, A.; Scudiero, D.; Skehan, P.; Shoemaker, R.; Paull, K.; Vistica, D.; Hose, C.; Langley, J.; Cronise, P.; Vaigro-Wolff, A. Feasibility of a high flux anticancer drug screen



using a diverse panel of cultured human tumor cell lines. *J Natl. Canc. Inst* **1991**, *83*, 757–766.

- [29] Boyd, M. R. The NCI In Vitro Anticancer Drug Discovery Screen. In *Anticancer Drug Development Guide: Preclinical Screening, Clinical Trials and Approval*; Teicher, B., Ed.; Humana Press: Totowa, New Jersey, 1995; pp 23–41.
- [30] Boyd, M.; Paull, K. D. Some practical considerations and applications of the National Cancer Institute in vitro anticancer drug discovery screen. *Drug Devel. Res* **1995**, *34*, 91–109.
- [31] Shi, L. M.; Fan, Y.; Myers, T. G.; Waltham, M.; Paull, K. D.; Weinstein, J. N. Mining the anticancer activity database generated by the U. S. National Cancer Institute's drug discovery program using statistical and artificial intelligence techniques. *Modeling and Scientific Computing* **1998**, *38*, 189–196.
- [32] Shi, L. M.; Myers, T. G.; Fan, Y.; O'Connor, P. M.; Paull, K. D.; Friend, S. H.; Weinstein, J. N. Mining the National Cancer Institute anticancer drug discovery database: cluster analysis of ellipticine analogs with p53-inverse and central nervous system-selective patterns of activity. *Molecular Pharmacology* **1998**, *53*, 241–251.
- [33] Shi, L. M.; Fan, Y.; Myers, T. G.; O'Connor, P. M.; Paull, K. D.; Friend, S. H.; Weinstein, J. N. Mining the NCI anticancer drug discovery databases: genetic function approximation for the QSAR study of anticancer ellipticine analogues. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 189–199.
- [34] O'Connor, P. M.; Jackman, J.; Bae, I.; Myers, T. G.; Fan, S.; Mutoh, M.; Scudiero, D. A.; Monks, A.; Sausville, E. A.; Weinstein, J. N.; Friend, S.; Fornace, A. J. J.; Kohn, K. W. Characterization of the p53 tumor suppressor pathway in cell lines of the National Cancer Institute anticancer drug screen and correlations with the growth-inhibitory potency of 123 anticancer agents. *Cancer Research* **1997**, *57*, 4285–4300.
- [35] Keskin, O.; Bahar, I.; Jernigan, R. L.; Beutler, J. A.; Shoemaker, R. H.; Sausville, E. A.; Covell, D. G. Characterization of Anticancer Agents by Their Growth-Inhibitory Activity

- and Relationships to Mechanism of Action and Structure. *Anticancer Drug Discovery* **2000**, *15*, 79–98.
- [36] Giuliani, A.; Colosimo, A.; Benigni, R.; Zbilut, J. On the constructive role of noise in spatial systems. *Physics Letters A* **1998**, *247*, 47–52.
- [37] Berry, M. W.; Dumais, S. T.; O'Brien, G. W. Using linear algebra for intelligent information retrieval. *SIAM Rev.* **1995**, *37*, 573–595.
- [38] Implementation of the self-organizing map method is a modification and extension of the SOM toolbox programming team (<http://www.cis.hut.fi/projects/somtoolbox/>) Matlab package and SOM\_PAK C code ([http://www.cis.hut.fi/research/som\\_pak/](http://www.cis.hut.fi/research/som_pak/)).
- [39] vanOsdol, W. W.; Myers, T. G.; Paull, K. D.; Kohn, K. W.; Weinstein, J. N. Use of Kohonen self-organizing maps to study the mechanism of action of chemotherapeutic agents. *Journal of the National Cancer Institute* **1994**, *86*, 1853–1859.
- [40] vanOsdol, W. W.; Myers, T. G.; Weinstein, J. N. Neural network techniques for informatics of cancer drug discovery. *Methods Enzymol* **2000**, *321*, 369–395.
- [41] Martin, Y. C.; Willet, P. *Designing Bioactive Molecules*. American Chemical Society: Washington D.C., 1998.
- [42] Pearlman, R. S.; Smith, K. M. Novel software tools for chemical diversity. *Perspectives in drug discovery* **1998**, *9-11*, 339–353.
- [43] Maggiora, G.; Johnson, M. A. *Concepts and Applications of Molecular Similarity*. John Wiley: New York, NY, 1990.
- [44] Randić, M. On characterization of chemical structure. *Journal of Chemical Information and Computer Sciences* **1997**, *37*, 672–687.
- [45] Burden, F. R. Molecular identification number for substructure searches. *Journal of Chemical Information and Computer Sciences* **1989**, *29*, 225–227.
- [46] Ihlenfeldt, W.-D. *personal communication*.

- [47] Manatunga, A. K.; Chen, S. Measuring cluster similarity across methods: sample size estimation for survival outcomes in cluster-randomized studies with small cluster sizes. Cubic clustering criterion. *Biometrics* **2000**, *56*, 616–621.
- [48] Christman, M. C.; Pontius, J. S. Bootstrap confidence intervals for adaptive cluster sampling. *Biometrics* **2000**, *56*, 503–510.
- [49] Hrach, K. Comparison of survival between two groups using software SAS, S-PLUS and STATISTICA. *Journal of Medical Informatics* **1997**, *45*, 31–33.
- [50] Kos, A. J.; Psenicka, C. Measuring cluster similarity across methods. *Psychol Rep* **2000**, *86*, 858–862.
- [51] Institute, S. Cubic clustering criterion. *Technical Report A-108* **1992**, pp 1–17.
- [52] Galitski, T.; Saldanha, A. J.; Styles, C. A.; Lander, E. J.; Fink, G. R. Ploidy regulation of gene expression. *Science* **1999**, *285*, 251–254.
- [53] Boyd, M. R. Status of the NCI preclinical antitumor drug discovery screen. In *Cancer: Principles and Practice of Oncology Updates*; DeVita, V. T., Hellman, S., Rosenberg, S. A., Eds.; Lippincott: Philadelphia, PA, 1989; pp 1–12 Vol. 3 No. 10.
- [54] Paull, K. D.; Hamel, E.; Malspeis, L. Prediction of biochemical mechanism of action from the in vitro antitumor screen of the National Cancer Institute. In *Cancer Chemotherapeutic Agents*; Foye, W. O., Ed.; American Chemical Society: Washington D.C., 1995; pp 9–46.
- [55] Koutsoukos, A. D.; Rubenstein, L. V.; Faraggi, D.; Simon, R. M.; Kalyandrug, S.; Weinstein, J. N.; Kohn, K. W.; Paull, K. D. Discrimination techniques applied to the NCI *in vitro* anti-tumor drug screen: predicting biochemical mechanism of action. *Stat. Med.* **1994**, *13*, 719–730.
- [56] Elion, G. B.; Hitchings, G. H. Metabolic basis for the actions of analogs of purines and pyrimidines. *Adv. Chemother.* **1965**, *2*, 91–177.

- [57] Chabner, B. A.; Longo, D. L. *Cancer Chemotherapy and Biotherapy: Principles and Practice*. Lippencot-Raven: Philadelphia and New York, 1996.
- [58] Hatse, S.; De Clercq, E.; Balzarini, J. Role of antimetabolites of purine and pyrimidine nucleotide metabolism in tumor cell differentiation. *Biochemical Pharmacology* **1999**, *58*, 539–555.
- [59] Izquierdo, M. A.; Shoemaker, R. H.; Flens, M. J.; Scheffer, G. L.; Wu, L.; Prather, T. R.; Scheper, R. J. Overlapping phenotypes of multidrug resistance among panels of human cancer-cell lines. *Int. J. Cancer* **1996**, *65*, 230–237.
- [60] Gharehbaghi, K.; Sreenath, A.; Hao, Z.; Paull, K. D.; Szekeres, T.; Cooney, D. A.; Krohn, K.; Jayaram, H. N. Comparison of biochemical parameters of benzamide riboside, a new inhibitor of IMP dehydrogenase, with tiazofurin and selenazofurin. *Biochem. Pharmacol.* **1994**, *48*, 1413–1419.
- [61] Kodama, E. N.; McCaffrey, R. P.; Yusa, K.; Mitsuya, H. Antileukemic activity and mechanism of action of cordycepin against terminal deoxynucleotidyl transferase-positive (TdT+) leukemic cells. *Biochem Pharmacol* **2000**, *59*, 273–281.
- [62] Barchi, J. J. J.; Cooney, D. A.; Ahluwalia, G. S.; Gharehbaghi, K.; Covey, J. M.; Hochman, I.; Paull, K. D.; Jayaram, H. N. Studies on the mechanism of action of 1-b-D-arabinofuranosyl-5-azacytosine (fazarabine) in mammalian lymphoblasts. *J. Exp. Ther. Oncol.* **1996**, *1*, 191–203.
- [63] Stryer, L. *Biochemistry*. W. H. Freeman: New York, NY, 1999.
- [64] Kim, H. K.; Nam, J. Y.; Han, M. Y.; Lee, E. K.; Choi, J. D.; Bok, S. H.; Kwon, B. M. Actinomycin D as a novel SH2 domain ligand inhibits Shc/Grb2 interaction in B104-1-1 (neu\*-transformed NIH3T3) and SAA (hEGFR-overexpressed NIH3T3) cells. *FEBS Lett* **1999**, *453*, 174–178.
- [65] Gedik, C. M.; Collins, A. R. Comparison of effects of fostriecin, novobiocin, and camptothecin, inhibitors of DNA topoisomerases, on DNA replication and repair in human cells. *Nucleic Acids Res.* **1990**, *18*, 1007–1013.

- [66] Cheng, A.; Balczon, R.; Zuo, Z.; Koons, J. S.; Walsh, A. H.; Honkanen, R. E. Fostriecin-mediated G2-M-phase growth arrest correlates with abnormal centrosome replication, the formation of aberrant mitotic spindles, and the inhibition of serine/threonine protein phosphatase activity. *Cancer Res.* **2000**, *58*, 3611–3619.
- [67] Zolnierowicz, S. Type 2A protein phosphatase, the complex regulator of numerous signaling pathways. *Biochem Pharmacol* **2000**, *60*, 1225–1235.
- [68] Frosina, G.; Rossi, O. Effect of topoisomerase poisoning by antitumor drugs VM 26, fostriecin and camptothecin on DNA repair replication by mammalian cell extracts. *Carcinogenesis* **1992**, *13*, 1371–1377.
- [69] Kawada, M.; Amemiya, M.; Ishizuka, M.; Takeuchi, T. Cytostatin, an inhibitor of cell adhesion to extracellular matrix, selectively inhibits protein phosphatase 2A. *Biochim Biophys Acta* **1999**, *1452*, 209–217.
- [70] Muller, K. Current status and recent developments in anthracenone antipsoriatics. *Curr Pharm Des* **2000**, *6*, 901–918.
- [71] Numann, J. F. Histamine increases  $[Ca^{2+}]_{in}$  and activates Ca-K and nonselective cation currents in cultured human capillary endothelial cells. *J Membr Biol* **2000**, *173*, 107–116.
- [72] Kuo, C. C.; Huang, R. C.; Lou, B. S. Inhibition of  $Na^{+}$  current by diphenhydramine and other diphenyl compounds: molecular determinants of selective binding to the inactivated channels. *Mol Pharmacol* **2000**, *57*, 135–143.
- [73] Zhou, Z.; Vorperian, V. R.; Gong, Q.; Zhang, S.; January, C. T. Block of HERG potassium channels by the antihistamine astemizole and its metabolites desmethylastemizole and norastemizole. *J Cardiovasc Electrophysiol* **1999**, *10*, 836–843.
- [74] Lee, M. G.; Xu, X.; Zeng, W.; Diaz, J.; Kuo, T. H.; Wuytack, F.; Racymaekers, L.; Muallem, S. Polarized expression of  $Ca^{2+}$  pumps in pancreatic and salivary gland cells. Role in initiation and propagation of  $[Ca^{2+}]_i$  waves. *J Biol Chem* **1997**, *272*, 15771–15776.

- [75] Hobman, T. C.; Woodward, L.; Farquhar, M. G. The rubella virus E1 glycoprotein is arrested in a novel post-ER, pre-Golgi compartment. *J Cell Biol* **1992**, *118*, 795–811.
- [76] Duncan, K. L.; Duncan, M. D.; Alley, M. C.; Sausville, E. A. Cucurbitacin E-induced disruption of the actin and vimentin cytoskeleton in prostate carcinoma cells. *Biochem Pharmacol*. **1996**, *52*, 1553–1560.
- [77] Bai, R.; Verdier-Pinard, P.; Gangwar, S.; Stessman, C. C.; McClure, K. J.; Sausville, E. A.; Pettit, G. R.; Bates, R. B.; Hamel, E. Dolastatin 11, a marine depsipeptide, arrests cells at cytokinesis and induces hyperpolymerization of purified actin. *Mol Pharmacol* **2001**, *59*, 462–469.
- [78] Bubb, M. R.; Senderowicz, A. M.; Sausville, E. A.; Duncan, K. L.; Korn, E. D. Jasplakinolide, a cytotoxic natural product, induces actin polymerization and competitively inhibits the binding of phalloidin to F-actin. *J Biol Chem* **1994**, *269*, 14869–14871.
- [79] Spector, I.; Braet, F.; Shochet, N. R.; Bubb, M. R. New anti-actin drugs in the study of the organization and function of the actin cytoskeleton. *Microsc Res Tech* **1999**, *47*, 18–37.
- [80] Brown, B. K.; Song, W. The actin cytoskeleton is required for the trafficking of the b cell antigen receptor to the late endosomes. *Traffic* **2000**, *2*, 414–417.
- [81] Rosado, J. A.; Sage, S. O. Activation of store-mediated calcium entry by secretion-like coupling between the inositol 1,4,5-trisphosphate receptor type II and human transient receptor potential (hTrp1) channels in human platelets. *Biochem J* **2001**, *356*(Pt 1), 191–198.
- [82] Koya, R. C.; Fujita, H.; Shimizu, S.; Ohtsu, M.; Takimoto, M.; Tsujimoto, Y.; Kuzumaki, N. Gelsolin inhibits apoptosis by blocking mitochondrial membrane potential loss and cytochrome c release. *J Biol Chem* **2000**, *275*, 15343–15349.
- [83] Cole, J. A. Down-regulation of protein kinase C by parathyroid hormone and mezerein differentially modulates cAMP production and phosphate transport in opossum kidney cells. *J Bone Miner Res* **1997**, *12*, 1223–1230.

- [84] Yoshida, M.; Yokokura, H.; Hidaka, H.; Ikekawa, T.; Saijo, N. Mechanism of antitumor action of PKC activator, gnidimacrin. *Int J Cancer* **1998**, *77*, 243–250.
- [85] Moreno, O. A.; Kishi, Y. Total synthesis and stereochemistry of cytoblastin. *Bioorg Med Chem* **1998**, *6*, 1223–1254.
- [86] de Jong, R. S.; Mulder, N. H.; Uges, D. R.; Sleijfer, D. T.; Hoppener, F. J.; Groen, H. J.; Willemse, P. H.; van der Graaf, W. T.; de Vries, E. G. Phase I and pharmacokinetic study of the topoisomerase II catalytic inhibitor fostriecin. *Br J Cancer* **1999**, *79*, 882–887.
- [87] Ruiter, G. A.; Zerp, S. F.; Bartelink, H.; van Blitterswijk, W. J.; Verheij, M. Alkyllysophospholipids activate the SAPK/JNK pathway and enhance radiation-induced apoptosis. *Cancer Res* **1999**, *59*, 2456–2463.
- [88] Grunicke, H. H.; Maly, K.; Tinhofer, I.; Giselsbrecht, S.; Kampfer, S.; Baier, G.; Uberall, F. Inhibition of phospholipase C and protein kinase C by alkylphosphocholines. *Drugs Today* **1998**, *34 Suppl F*, 3–14.
- [89] Kamishohara, M.; Kenney, S.; Domergue, R.; Vistica, D. T.; Sausville, E. A. Selective accumulation of the endoplasmic reticulum-golgi intermediate compartment induced by the antitumor drug KRN5500. *Exp. Cell Res.* **2000**, *256*, 468–479.
- [90] Rennecke, J.; Richter, K. H.; Haussermann, S.; Stempka, L.; Strand, S.; Stohr, M.; Marks, F. Biphasic effect of protein kinase C activators on spontaneous and glucocorticoid-induced apoptosis in primary mouse thymocytes. *Biochim. Biophys. Acta* **2000**, *1497*, 289–296.
- [91] Clarke, H.; Ginanni, N.; Laughlin, K. V.; Smith, J. B.; Pettit, G. R.; Mullin, J. M. The transient increase of tight junction permeability induced by bryostatin 1 correlates with rapid downregulation Of protein kinase C- $\alpha$ . *Exp. Cell Res.* **2000**, *261*, 239–249.
- [92] Sidovar, M. F.; Kozlowski, P.; Lee, J. W.; Collins, M. A.; He, Y.; Graves, L. M. Phosphorylation of serine 43 is not required for inhibition of c-Raf kinase by the cAMP-dependent protein kinase. *J. Biol. Chem.* **2000**, *275*, 28688–28694.

- [93] Tedeschi, A.; Lorini, M.; Galbiati, S.; Gibelli, S.; Miadonna, A. Inhibition of basophil histamine release by tyrosine kinase and phosphatidylinositol 3-kinase inhibitors. *Int. J. Immunopharmacol.* **2000**, *22*, 797–808.
- [94] Choi, Y. H.; Baek, J. H.; Yoo, M. A.; Chung, H. Y.; Kim, N. D.; Kim, K. W. Induction of apoptosis by ursolic acid through activation of caspases and down-regulation of c-IAPs in human prostate epithelial cells. *Int. J. Oncol.* **2000**, *17*, 565–571.
- [95] Wang, X. F.; Chan, H. C. Adenosine triphosphate induces inhibition of Na(+) absorption in mouse endometrial epithelium: A Ca(2+)-dependent mechanism. *Biol Reprod* **2000**, *63*, 1918–1924.
- [96] Turchi, L.; Loubat, A.; Rochet, N.; Rossi, B.; Ponzio, G. Evidence for a direct correlation between c-Jun NH2 Terminal Kinase 1 activation, cyclin D2 expression, and G(1)/S phase transition in the murine hybridoma 7TD1 cells. *Exp Cell Res* **2000**, *261*, 220–228.
- [97] Ecay, T. W.; Dickson, J. L.; Conner, T. D. Wortmannin inhibition of forskolin-stimulated chloride secretion by T84 cells. *Biochim Biophys Acta* **2000**, *1467*, 54–64.
- [98] Yamashita, Y.; Ho, J. J.; Farrelly, E. R.; Hirakawa, K.; Sowa, M.; Kim, Y. S. Forskolin and phorbol ester have opposite effects on the expression of mucin-associated sialyl-Lewis(a) in pancreatic cancer cells. *Eur J Cancer* **2000**, *36*, 113–120.
- [99] Rabow, A. A.; Covell, D. G. *manuscript in preparation*.
- [100] Aoki, M.; Blazek, E.; Vogt, P. K. A role of the kinase mTOR in cellular transformation induced by the oncoproteins P3k and Akt. *Proc Natl Acad Sci U S A* **2001**, *98*, 136–141.
- [101] Kim, J. E.; Chen, J. Cytoplasmic-nuclear shuttling of FKBP12-rapamycin-associated protein is involved in rapamycin-sensitive signaling and translation initiation. *Proc Natl Acad Sci U S A* **2000**, *97*, 14340–14345.
- [102] Kubo, A.; Nakagawa, K.; Varma, R. K.; Conrad, N. K.; Cheng, J. Q.; Lee, W. C.; Testa, J. R.; Johnson, B. E.; Kaye, F. J.; Kelley, M. J. The p16 status of tumor cell lines



- identifies small molecule inhibitors specific for cyclin-dependent kinase 4. *Clin Cancer Res* **1999**, *5*, 4279–4286.
- [103] MacCarthy-Morrogh, L.; Townsend, P. A.; Purohit, A.; Hejaz, H. A. M.; Potter, B. V.; Reed, M. J.; Packham, G. Differential effects of estrone and estrone-3-O-sulfamate derivatives on mitotic arrest, apoptosis, and microtubule assembly in human breast cancer cells. *Cancer Res* **2000**, *60*, 5441–5450.
- [104] Mathews, C. K.; van Holde, K. E.; Ahern, K. G. *Biochemistry*. Addison Wesley Longman: New York, NY, 2000.
- [105] Morre, D. J.; Grieco, P. A.; Morre, D. M. Mode of action of the anticancer quassinoids— inhibition of the plasma membrane NADH oxidase. *Life Sci* **1998**, *63*, 595–604.
- [106] Piero, C.; Vallejo, S.; Cercas, E.; Llergo, J. L.; Lafuente, N.; Matesanz, N.; Rodriguez-Manas, L.; Sanchez-Ferrer, C. Thapsigargin induces apoptosis in cultured human aortic smooth muscle cells. *J. Cardiovasc. Pharmacol* **2000**, *36*, 676–680.
- [107] Shifrin, V. I.; Anderson, P. K. Trichothecene mycotoxins trigger a ribotoxic stress response that activates c-Jun N-terminal kinase and p38 mitogen-activated protein kinase and induces apoptosis. *J Biol Chem* **1999**, *274*, 13985–13992.
- [108] Nakashima, T.; Miura, M.; Hara, M. Tetrocarcin A inhibits mitochondrial functions of Bcl-2 and suppresses its anti-apoptotic activity. *Cancer Res* **2000**, *60*, 1299–1235.
- [109] Chung, K. C.; Kim, S. M.; Rhang, S.; Lau, L. F.; Gomes, I.; Ahn, Y. S. Expression of immediate early gene pip92 during anisomycin-induced cell death is mediated by the JNK- and p38-dependent activation of Elk1. *Eur J Biochem* **2000**, *267*, 4676–4684.
- [110] Alizadeh, A. A.; Eisen, M. B.; Davis, R. E.; Ma, C.; Lossos, I. S.; Rosenwald, A.; Boldrick, J. C.; Sabet, H.; Tran, T.; Yu, X.; Powell, J. I.; Yang, L.; Marti, G. E.; Moore, T.; Hudson, J.; Lu, L.; Lewis, D. B.; Tibshirani, R.; Sherlock, G.; Chan, W. C.; Greiner, T. C.; Weisenburger, D. E.; Armitage, J. O.; Warnke, R.; Levy, R.; Wilson, W.; Grever, M. R.; Byrd, J. C.; Botstein, D.; Brown, P. O.; Staudt, L. M. Distinct types of

diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **2000**, *403*, 503–511.

- [111] Perou, C. M.; Jeffrey, S. S.; , van de Rijn, M.; Rees, C. A.; Eisen, M. B.; Ross, D. T.; Pergamenchikov, A.; Williams, C. F.; Zhu, S. X.; Lee, J. C. F.; Lashkari, D.; Shalon, S.; Brown, P. O.; Botstein, D. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. (USA)* **1999**, *96*, 9212–9217.
- [112] Perou, C. M.; Sorlie, T.; Eisen, M. B.; van der Rijn, M.; Jeffrey, S. S.; Rees, C. A.; Pollack, J. R.; Ross, D. T.; Johnsen, H.; Akslen, L. A.; Fluge, O.; Pergamenschikov, A.; Williams, C.; Zhu, S. X.; Lonning, P. E.; , Borresen-Dale, A.-L.; Brown, P. O.; Botstein, D. Molecular portraits of human breast tumours. *Nature* **2000**, *406*, 747–752.
- [113] Ross, D. T.; Scherf, U.; Eisen, M. B.; Perou, C. M.; Rees, C.; Spellman, P.; Iyer, V.; Jeffrey, S. S.; van de Rijn, M.; Waltham, M.; Pergamenschikov, A.; Lee, J. C. F.; Laskhari, D.; Shalon, D.; Myers, T. G.; Weinstein, J. N.; Botstein, D.; Brown, P. O. Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics* **2000**, *24*, 227–235.
- [114] Saitoh, F.; Hiraishi, K.; Adachi, M.; Hozumi, M. Induction by 5-aza-2'-deoxycytidine, an inhibitor of DNA methylation of Le<sup>y</sup> antigen, apoptosis and differentiation in human lung cancer cells. *Anticancer Research* **1995**, *15*, 2137–2144.
- [115] Sokoloski, J. A.; Beardsley, G.; Sartorelli, A. C. Mechanism of the induction of the differentiation of HL-60 cells by antifolates. *Cancer Communication* **1989**, *1*, 199–207.

**Figure 1:**

**Panel A:** The conditioned growth inhibition ( $\log(GI_{50})$ ) data consists of an  $M \times N$  matrix of data elements. In the example above there are 533 data vectors ( $M = 533$ ) for the ExMOA data set (an extension of the 122 MOA set containing 362 compounds see text.) There are 80 components for each data vector measuring the response across the different cell lines ( $N = 80$ ). The data for two of the 80 dimensions for the 533 data vectors are shown as blue dots. A set of  $P$  cluster vectors are chosen to represent the data space. The number of cluster vectors and the map dimensions are selected according to the first two principle components found with single value decomposition (SVD) analysis of this data. These cluster vectors are shown as open red circles (in the example above  $P = 153$ .) Shown here are data for only 2 of the 80 dimensions in our data vectors.

**Panel B:** SOM clustering of the initial coordinates shown in Panel A. The SOM cluster vectors minimizes the distances between data(blue dots) and cluster vectors(red circles).

**Panel C:** To make the information contained in the high dimensional clustering space accessible for drug discovery, the  $P$  clusters in  $N$  dimensions are projected on to a two dimensional map. This mapping uses a non-linear function to transform the data such that each cluster vector is uniformly represented in the two dimensional map. This reorganization stretches the data space such that the map has a finer discrimination where more data is present.

**Panel D:** SOM cluster map for the ExMOA data set (see text). The standard anticancer agents are clustered on to a 17x9 hexagonal array. The hexagons at each node position correspond to the number of agents clustered at these loci. Cluster distances are indicated by the colors between each node: close and far neighbors are separated by dark and light blue colors, respectively. A horizontal line below row six indicates separation between agents that act to disrupt mitosis from agents that act by inhibiting nucleic acid biosynthesis. Located in map margins are 2-D structures of representative compounds in selected regions. For example, the camptothecins are all located in the lower right portion of this SOM map. More information on understanding the SOM map can be found in a short primer contained in the supporting information section, and in the 3d Mind manual pages at <http://spheroid.ncifcrf.gov>.

**Figure 2:** Sample of structure/function (S/F) correlation for different forms of data conditioning. Panel A: Z-score, capping and no alteration for unknown data(NaN). Panel B: Z-score, no capping and no alteration for unknown data. Panel C: Z-score, no capping and replacement of unknowns with group average (mean). Best cases occur for Z-score, capping and no alteration for unknown data. The S/F correlation coefficient,  $\rho$ , is shown in each panel.

**Figure 3:** Structure/function (S/F) correlation versus ratio of map dimensions. The total number of clusters was kept as near a possible to 153. Maximum average S/F correlations occur for a ratio of map dimension of 1.89. This corresponds to the SOM map dimensions of 17x9. The filled circles represent averages of correlation coefficient with the bars at the 25<sup>th</sup> and 75<sup>th</sup> percentiles. The open circles and dashed line represent the base case found for each map ratio. The conventions for filled circles, bars, open circles, and dashed lines are used throughout Figures 3 through 6.

**Figure 4:** Structure/function (S/F) correlation versus number of clusters. Repeat SOM maps were generated for different cluster numbers. The solid line represents an exponential function fitting of the data. The 99 percent asymptote occurs at 110 clusters (i.e. this level captures 99% of the possible S/F correlation.)

**Figure 5:** Structure/function (S/F) correlation versus number of cell lines. Highest correlations occur for the greatest number of cell lines. Plateaus are observed in the average S/F correlation for 50-60 cell lines and 20-30 cell lines. Fewer than 20 cell lines drastically reduce the S/F correlation coefficient.

**Figure 6:** Structure/function (S/F) correlation versus data completeness (Panel A) and versus Z-score threshold (Panel B). Incomplete data sets are reasonably well tolerated above 60 percent. When greater than 40 percent of the data is removed, the S/F correlation declines continuously. Lower panel displays correspondence between S/F correlation coefficient and Z-score. Low Z-scores indicate a relatively flat cellular response pattern (low signal strength.)

**Figure 7:** Complete DTP map for the 20K compounds tested in the NCI's tumor cell screen.

Map consists of 41x26 clusters. Color bar at lower right indicates distance between clusters (red:close, black:intermediate, purple:far). Fifty regions have been defined on this map that group together individual clusters with the most similar response profiles. These regions are assigned to six functional categories according to their apparent cellular activity: mitosis(M), nucleic acid synthesis(S), membrane transport and integrity(N), phosphatase and kinase mediated cell cycle regulation(P). Regions Q and R have not been assigned to an activity class. See the text for details.

### Figure 8:

**Panel A:** The projections onto the complete SOM map of a set of 171 clinically evaluated anticancer agents. Shown on the map are the compound location (yellow hexagons), name and National Cancer Institute NSC identification number (given in parentheses) with abbreviation labels correspondence; Mercaptopurines (k34.17): thioguanine(752), 6-mercaptopurine(755), B-TGDR(71261), A-TGDR(71851) and ARA-6-MP(406021); Topo II Inhibitors(k41.26): oxanthrazole(349174), acodazole hydrochloride(305884), deoxydoxorubicin(267469), rubidazone(164011), doxorubicin/Adriamycin(123127), VM-26/teniposide(122819) and daunorubicin/daunomycin(82151); and, Bifunctional Alkylating Agents (k28.25): chlorambucil(3088), thio-tepa(6396), melphalan(8806), triethylenemelamine(9706), pipobroman(25154), uracil nitrogen mustard(34462), Yoshi-864(102627), dianhydrogalatitol(132313), piperazinedione(135758), AZQ(182986), teroxirone(296934) and hepsulfam(329680). The location of a compound is given by its map coordinates, for example, rhizoxin (NSC identification number 332598) is projected to the upper left at the loci k5.1 (5 rows down in the first column).

**Panel B:** Locations of compounds on the complete SOM map according to their cellular activity. Highlighted regions represent map clusters where assigned cellular activity is supported by literature references. Shown are a subset of molecular activity classes corresponding for a number of the M,N,P and S sub-regions (see text for details.) The size of the colored hexagons represents the number of data vectors belonging to that class found at that map coordinate, for example, the taxanes are found chiefly at k9.1 and k9.2 (large cyan hexagons), but a few are also found at k8.1 and k10.1 (small cyan hexagons.) The shown hexagons represent a subset

of both regions and classes contained in the SOM map. The complete map can be explored via the internet at <http://spheroid.ncifcrf.gov>.

**Figure 9:**

**Panel A:** Averaged intra and inter-panel pairwise correlation coefficients. Values along the diagonal represent average intra-panel pairwise correlation coefficients. For example, the SCL tumor cells exhibit the highest intra-panel correlation coefficient, while the LEU and COL tumor cells exhibit the lowest value. Values above the diagonal represent the averaged inter-panel pairwise correlation coefficients. For reference, the LEU:CNS and LNS:CNS panels are the most negatively and positively correlated tumor panels. The values above and below the diagonal are identical and are shown with two different scales for ease of interpretation.

**Panel B:** Histogram of pairwise correlation coefficients. Intra-panel correlations are shown in solid gray bars while inter-panel correlations are shown in open black bars. Mean and standard deviation values for the intra- and inter-panel distributions are  $0.18 \pm 0.16$  and  $-0.2 \pm 0.16$ , respectively.

**Figure 10:** Intra-panel averaged responses superimposed on the complete SOM map. The hexagon located at each of the 1066 map clusters displays tumor panel's average response to 20K compounds, colored, according to sensitive(red) and insensitive(blue) cellular activity; sized proportional to the magnitude of this activity.

**Figure 11:**

**Panel A:** Histograms of pairwise correlation coefficients obtained from random pairs of cellular responses and concentration pairs from the same compound. Duplicate concentration pairs are shown as open bars and the 'random pairs' are shown as closed gray bars. The leftmost open bar identifies the most similar pairs, those with a zero distance apart (in the same cluster). Cluster definitions are based on SOM analysis. There are 65 duplicate concentration pairs in this first bar. This bar is 5.0 standard deviation units away from the random pair mean of 9.9. The random pairs uses the scaling factor of 32.4, with the maximal number of pairs of 2236 at a distance of 11.1. There are 241 duplicate concentration pairs and 21,156 random pairs. The

lower plot (black bars) shows the fraction of pairs that contains a data vector with low signal strength (less than 0.15 units of absolute deviation). The fraction increases to the right with increasing distance likely indicating the decreasing reliability of the pairing.

**Panel B:** Histograms of pairwise correlation coefficients based on random (gray) and duplicate concentration pairs (open black). Correlation coefficients are obtained using the COMPARE program. The most and least similar pairs appear at the leftmost and rightmost portions of the graph, respectively. The random pairs uses the scaling factor of 15.7, with the maximal number of pairs of 1082 at a correlation coefficient of 0.258. There are 241 duplicate concentration pairs and an a sample of 11,210 random pairs. These random pairs are an unbiased sample of the complete set of random pairs (21,156 pairs). As above, the lower plot (black bars) shows the fraction of pairs that contains a data vector with low signal strength (less than 0.15 units of absolute deviation).

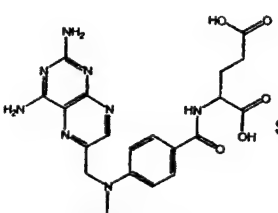
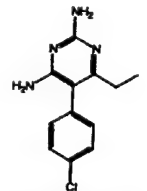
**Table I. Data Conditioning Structure/Function Correlation**

Missing Treatment	Normalization	Capping	S/F Corr. Mean	S/F Corr. Std. Dev	No. Map Samples
NaN	Raw	No Cap	0.7820	0.0648	20
NaN	Z-score	No Cap	0.9002	0.0182	40
NaN	<b>Z-score</b>	<b>Cap <math>\pm 3</math></b>	<b>0.9185</b>	<b>0.0147</b>	<b>40</b>
Mean	Z-score	Cap $\pm 3$	0.8654	0.0339	40

Multiple SOM maps were generated from random starting conditions for different combinations of data conditioning with respect to normalization(Raw vs. Z-score), capping (none vs.  $\pm 3$ ) and treatment of missing data (replace with mean vs. NaN:unknown). The correlation coefficient is determined between each of these SOM maps and the SOM clustering based on the structural descriptors. The basis of this comparison is that maps with the highest structure/function correlations are most desirable (the best treatment is shown in bold face.) Values represent averages for total number of samples.

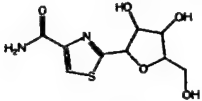
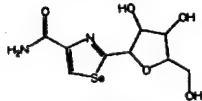
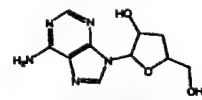
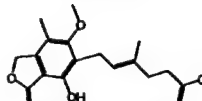
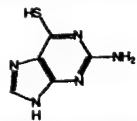


**Table II. Anti-folates**

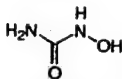
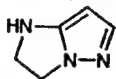
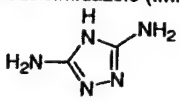
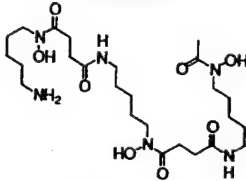
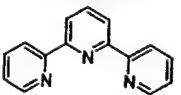
Structure	Region	Cluster Loci	NSC #
 Methotrexates	S4	k37.22 k38.22	740 680399
			3073 682306
			174121 687352
			607301 690436
			623017 694477
			626715 694478
			633713 694480
			654830 694481
			666783 694482
			667640 694483
			667641 694484
			669270 694485
			672140 694775
			672141 695788
			677942
 Pyremethamines	S4	k36.21 k37.21	3061 382034
			3062 382035
			7364 382044
			302325 382046
			319947 382049
			330465 382053
			372939 382054
			372944 602313
			372950 602314
			382032

**Table III. IMPDH**

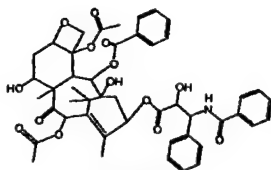
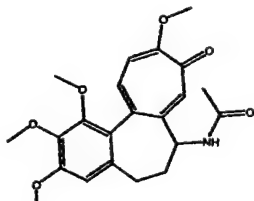
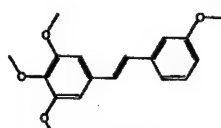
**Agents**

Structure	Region	Cluster Loci	NSC #
	Q5	k37.14 k38.13	286193
Tiazofurin			
	Q5	k38.13	340847
Selofurin			
	Q5	k37.14	63984
Cordycepin			
	S4	k37.20	129185
Mycophenolic acid			
	S6	k34.17	752 755 48388 71261 71851 406021 647471
6-mercaptopurine, 6-thioguanine			

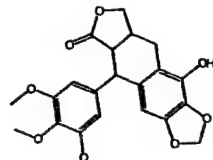
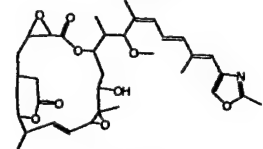
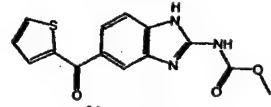
**Table IV. Ribonuclease  
Reductase Agents**

Structure	Cluster	
	Region	Loci
 Hydroxyurea	S3	k39.25
 Pyrazoloimidazole (IMPY)	S3	k39.25
 Guanazole	S3	k39.25
 Deferoxamine	S2	k32.22
 Terpyridine	S2	k31.23 k32.22
		3905 640499 640500 676944

**Table V. Mitotic Agents**

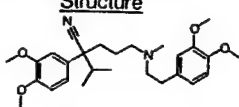
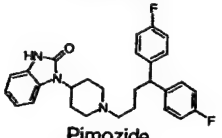
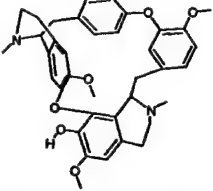
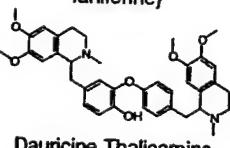
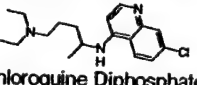
Structure	Cluster		NSC #
	Region	Loci	
 Taxanes	M2		125973 654374
			600220 656177
			600221 658831
			600222 661746
			600223 662158
		k9.1	608832 662159
		k9.2	628503 662160
			647752 662161
			647753 664401
			651195 664402
			651196 664403
			653244 etc.
			757 352277
			9170 373031
			33410 373032
 Colchicines	M2		33411 374979
			172946 374980
		k6.1	186301 376251
		k7.1	221662 612115
			249278 612116
			315260 618049
			320301 353494
			328403 354974
			335989 354975
			343493 etc.
 Trimethoxystilbenes	M1		638490 638486
		k3.1	638494 638411
		k3.2	638485 638390
		k3.3	638492 638404
		k5.2	638488 638499
			641484 638493
			638403 638493

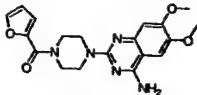
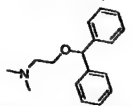
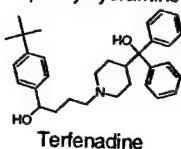
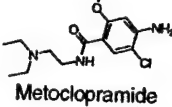
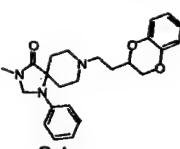
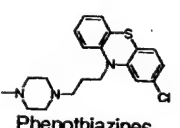
55

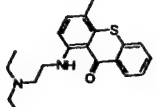
 Peltatins	M1	k1.2	24819
		k5.1	
		k5.2	
		k6.2	
 Rhizoxin	M1	k5.1	332598
 Nocodazole	M1	k1.1	238159

55

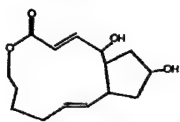
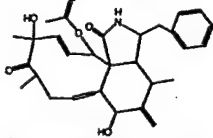
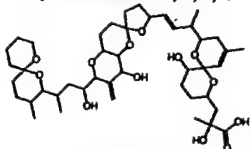
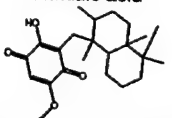
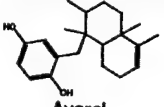
**Table VI. Channel Agents**

Structure	Region	Cluster Loci	NSC #
 Verapamil	N10	k9.8 k11.8	632821 657799
 Pimozide	N10	k9.7	170984
 Tetrandrines {fangchinoline, oxyacanthine, funiferine}	N10 P13	k6.6 k7.8 k8.7	77036 251534 77037 269189 91771 369310 93135 615580 93674 629742 97338 629744 105130 629745 105131 629746 135070 645315 189487
 Dauricine, Thallicarpine	N10	k8.7 k9.8 k10.8 k11.6	36413 68075 146267 209759
 Chloroquine Diphosphate	N10	k9.7	14050

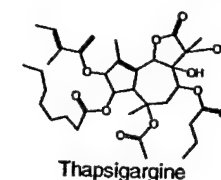
 Prazosin	P14	k7.4	292810
 Diphenylhydramine	N10	k11.8	665800
 Terfenadine	N10	k11.8	665802
 Metoclopramide	P11	k9.12	354467
 Spiperones	N11	k8.11	665340 665740 665742 665759 665771 665789 665862 665863 665873
 Phenothiazines {Prochlorperazine, Fluorphenothiazine, Trifluopromazine, Clopenthixol}	N10	k9.6 k9.8 k9.11 k11.8	17473 53638 64087 665801

 Lucanthone	N10	k9.6 k10.7	14574 20534
---	-----	---------------	----------------

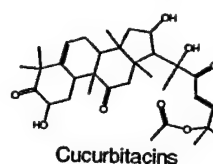
**Table VII. Golgi  
Disrupting Agents**

Structure	Region	Cluster Loci	NSC #
			56310
	N11	k12.10	89671
	P11	k9.16	656202
Brefeldin A			657326
			671928
	N11	k11.12	174119
	P7	k11.17	175151
	N1	k24.2	209835
Cytochalasins A,D,E,H	N3	k28.2	305222
	M2	k10.1	
	N7	k16.6	677083
Okadaic acid			
	N12	k17.8	311040
	N7	k19.5	647642
Limaquinone, Illimaquinone			
	N5	k21.5	306951
Avarol			

57



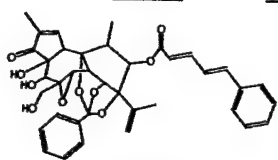
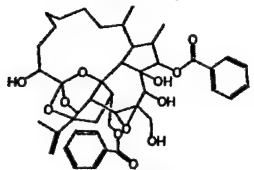
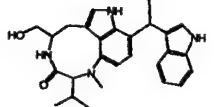
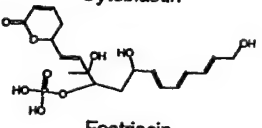
M2 K9.3 299933



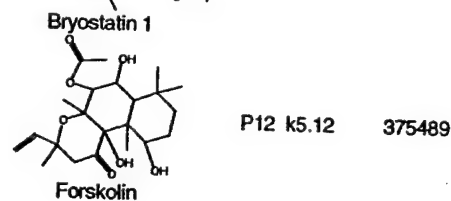
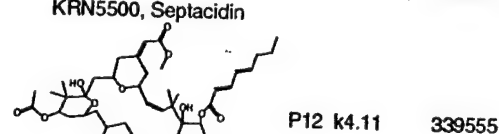
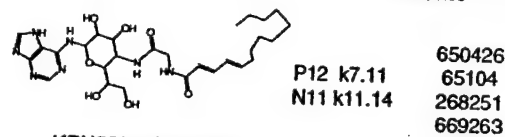
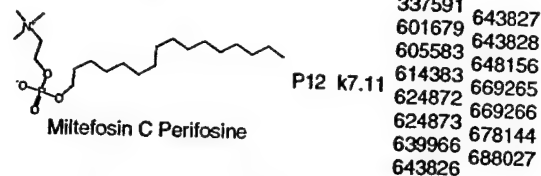
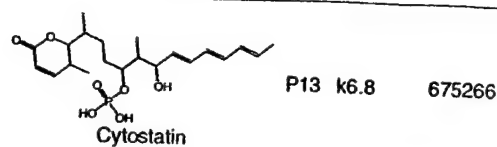
49451  
49452  
94743  
106399  
N11 k11.12 112166  
k11.13 112167  
144153  
308606  
521777

57

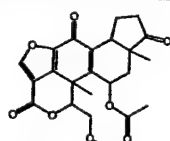
**Table VIII. Cell Cycle  
P Region Agents**

Structure	Cluster Region Loci	NSC #
	P13 k6.9	239072 266186 329507 339875 623310
Phorbols (Mezerein, Huratoxin, Prostratin)		
	P13 k6.9	252940
Gnidimacrin		
	P13 k6.9	654239
Cytoablastin		
	P13 k6.8	339638
Fostriecin		

58

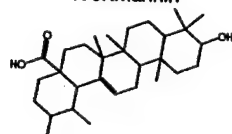


59



P12 k5.12 627609

Wortmannin

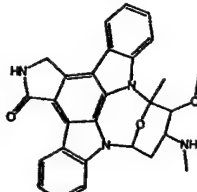
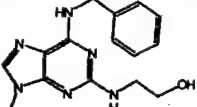
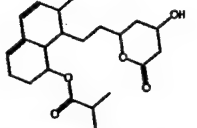
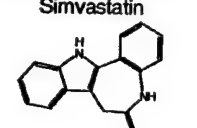


P12 k4.11 4060

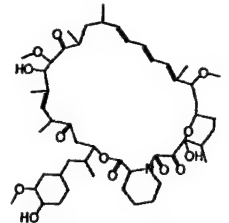
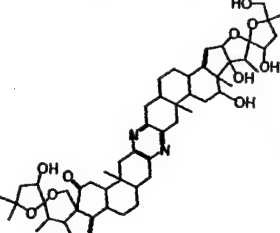
Ursolic acid

58

**Table IX. Cyclin Dependent  
Kinase Agents**

Structure	Cluster		
	Region	Loci	NSC #
	P2	k10.22 k10.23	618487 638850
Staurosporine, UCN-01			
	P1 R4	k18.26 k30.21	666096
Olomoucine			
	P2	k12.24	281245 633781 633782
Compactin, Lovastatin Simvastatin			
	P2	k11.24	672231
Paullone Analog			

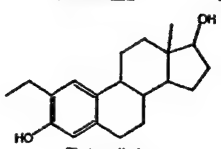
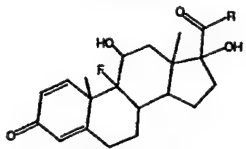
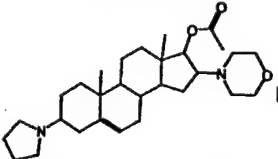
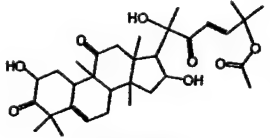
59

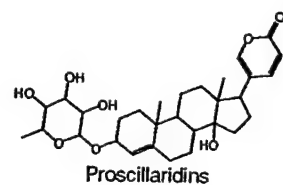
	P2	k11.26	226080 606698 606699 643248 683864
Rapamycin			
	P2	k14.26	363979 363980 363981 378727 378731 378732 378734 378735 378736
Cephalostatin			

59



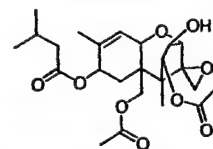
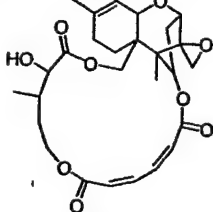
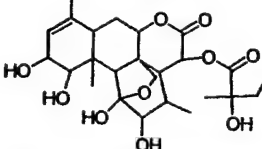
Table X. Steroids

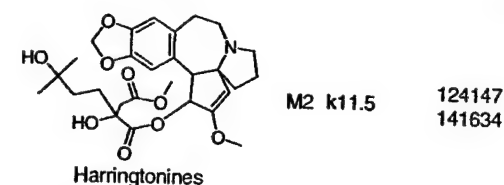
Structure	Region	Cluster Loci	NSC #
 Estradiols	M1	k1.4	667048 682429
			669229 682505
			669231 682597
			671042 683125
			673652 683688
			678473 684423
 Predisolones	P6 P11	k4.17 k8.15	R=CH <sub>2</sub> OH R=H or Ester k4.17 k8.15 12174 12600 13397 33001 34521 47438 63549 77021
 Cycloprotobuxine Analogs	N10	k8.7 k8.8	677955 677961 677956 677962 677957 677963 677959 689620 677960 689621
 Cucurbitacins	N11	k11.12 k11.13	49451 49452 94743 106399 112166 112167 144153 308606 521777



P9	k15.17 k16.17	7521 251698
		7525 345646
		7534 364373
		93134 619323
		123976 648338
		135036 650471
		135077 656598
		135687 664161
		135688 676514
		135689 677588
		143925 682561
		234669 688285
		243022 694454
		251692

**Table XI. M Region Antibiotics**

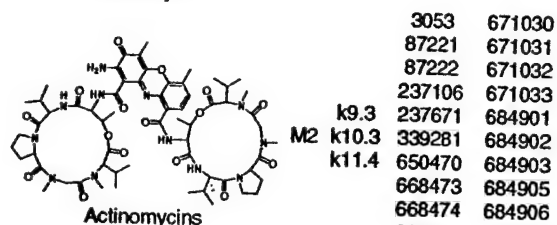
Structure	Region	Cluster Loci	NSC #
 Trichothecenes (T-2 Toxin, Anguidine, Scirpentriol, Ht-2 Toxin, Bruceoside E, F)	M2	k11.5	138780 656902
			141537 656903
			269142 656904
			278571 673352
			294913 673353
			656901 673354
 Verrucaric acid	M2	k11.5 k12.4	126728 291312
			269753 292463
			269754 327993
			269756 328166
			269757 328167
			269760 375726
			283445 604976
 Glaucarubin, Holacanthone	M2	k12.4	14975 341651
			126765 364170
			126765 364170
			132791 368671
			267709 688274
			279503 693539
			290494



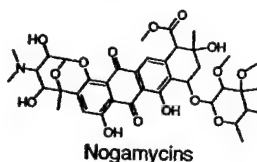
M2 k11.5 124147  
141634



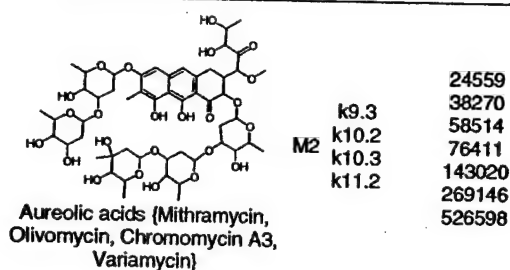
P11 k8.12 76712  
M2 k9.4 147340



3053 671030  
87221 671031  
87222 671032  
237106 671033  
M2 k9.3 237671 684901  
M2 k10.3 339281 684902  
k11.4 650470 684903  
668473 684905  
668474 684906  
668475 684907  
668476 684908



k1.11 70845 143491  
P12 k9.3 82151 265211  
M2 k9.4 86005 265450  
N11 k10.3 102815 267469  
k11.4 112929 639655  
k10.12 116555 670120



24559  
38270  
k9.3 58514  
M2 k10.2 76411  
k10.3 143020  
k11.2 269146  
526598

**Table XII. Replicate Concentration Pairs Test**

<b>3d Mind</b>			<b>COMPARE</b>		
Duplicate Pairs <sup>a</sup> (Euclid. Dist.)	Z-score <sup>c</sup>	Random Pairs <sup>b</sup>	Duplicate Pairs <sup>a</sup> (Corr. Coeff.)	Z-score <sup>c</sup>	Random Pairs <sup>b,d</sup>
25 (0.0)	5.0	24	25 (0.88)	3.0	0
50 (0.0)	5.0	24	50 (0.81)	2.6	77
75 (4.5)	2.7	24	75 (0.75)	2.3	279
100 (5.2)	2.4	323	100 (0.69)	2.0	621
125 (6.6)	1.7	1414	125 (0.64)	1.7	1038

<sup>a</sup>Duplicate pairs consist of all compounds from the Extended Mechanism of Action (ExMOA) set which have been measured at different maximum concentration levels.

<sup>b</sup>All pairings of compounds from the DTP Mechanism of Action set which differ in mechanism of action class. All data vectors have a signal strength greater than 0.08 absolute deviation units.

<sup>c</sup>  $Z_{score} = (\rho - \bar{\rho})/\sigma$  where  $\rho$  is the Euclidian distance (3d Mind) or the correlation coefficient (COMPARE),  $\bar{\rho}$  is the mean of the random pairs, and  $\sigma$  is the standard deviation of the random pairs. For example, a Z-score of 5.0 indicates five standard deviation units more similar than the mean for the random pairs. The value in the table is the Z-score of the last qualifying pair.

<sup>d</sup> The COMPARE random pairs have been normalized by sample size, normalizing ratio is 1.89, unnormalized random pairs 0, 41, 148, 329, 550 for duplicates 25, 50, 75, 100, and 125 respectively.

Note: Tables II-XI are generated from Microsoft Word files

Table II FOLS

Table III IMP

Table IV RR

Table V MIT

Table VI CH

Table VII GOL

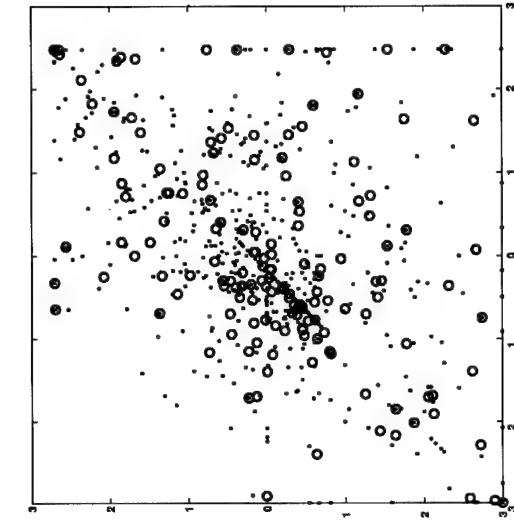
Table VIII PoKin

Table IX CDK

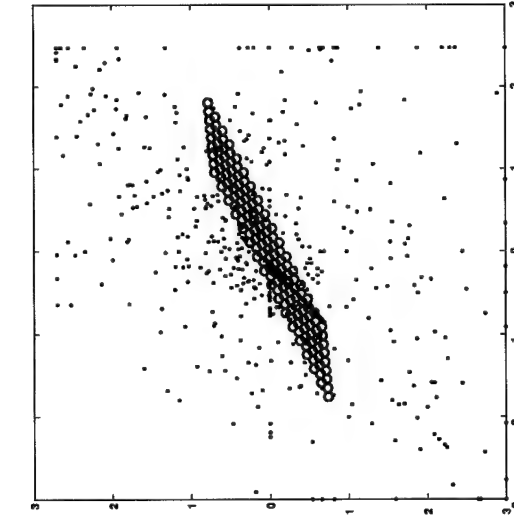
Table X STER

Table XI ANTIBOIS

**Fig. 1**

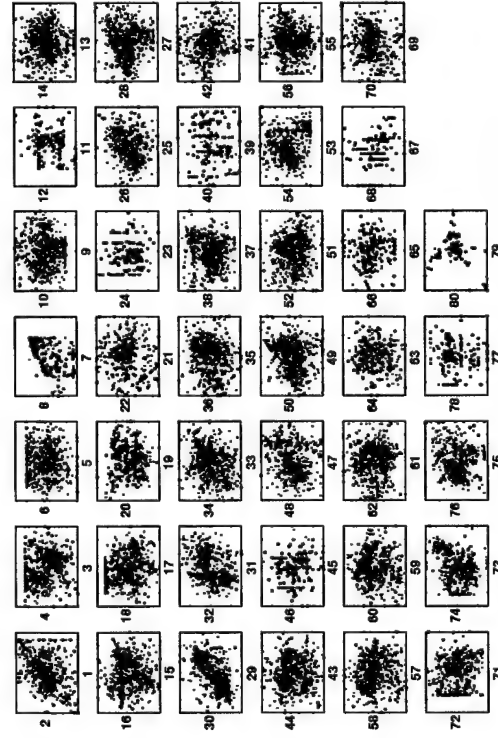


**SOM Clustering**  
 (2 of the 80 dimension shown)



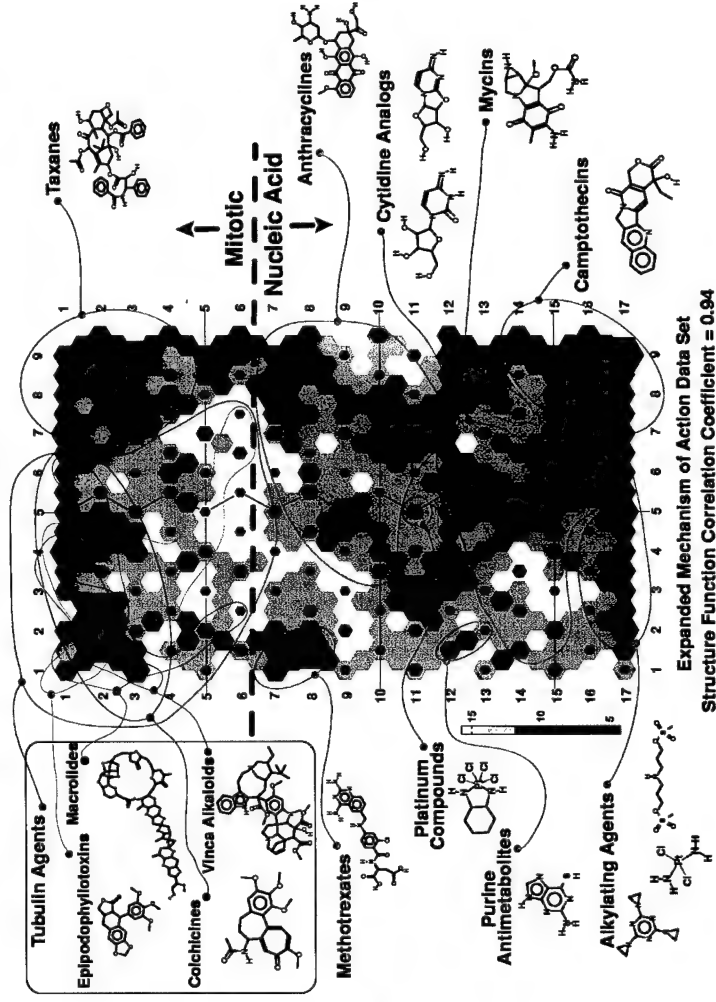
○ - Cluster Vector (2 of the 80 dimensions shown)  
 ● - Data Vector (2 of the 80 dimensions shown)

**C.**



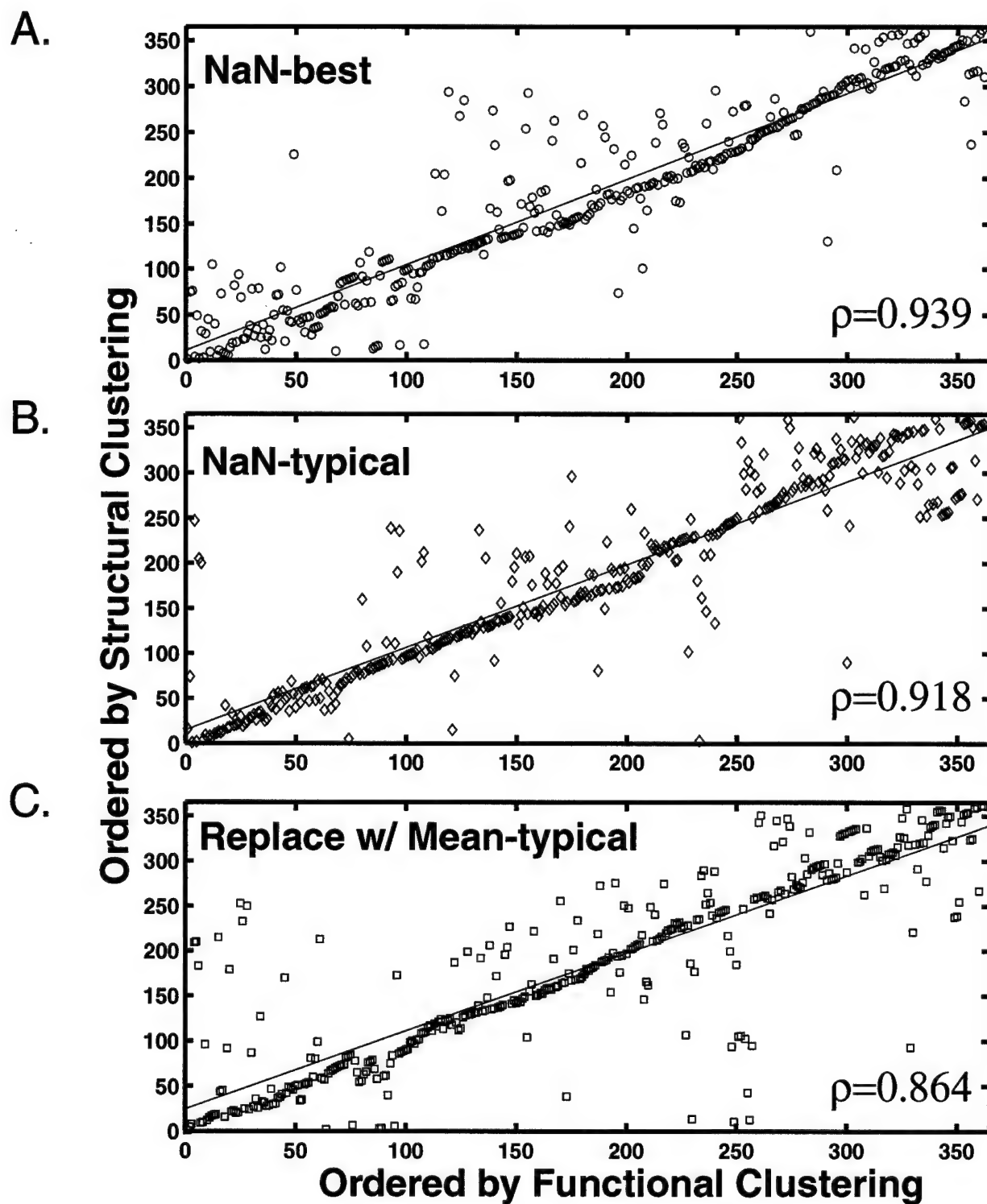
**D.**

# Anti-Cancer Biological Response Chemical Space



**Fig. 2**

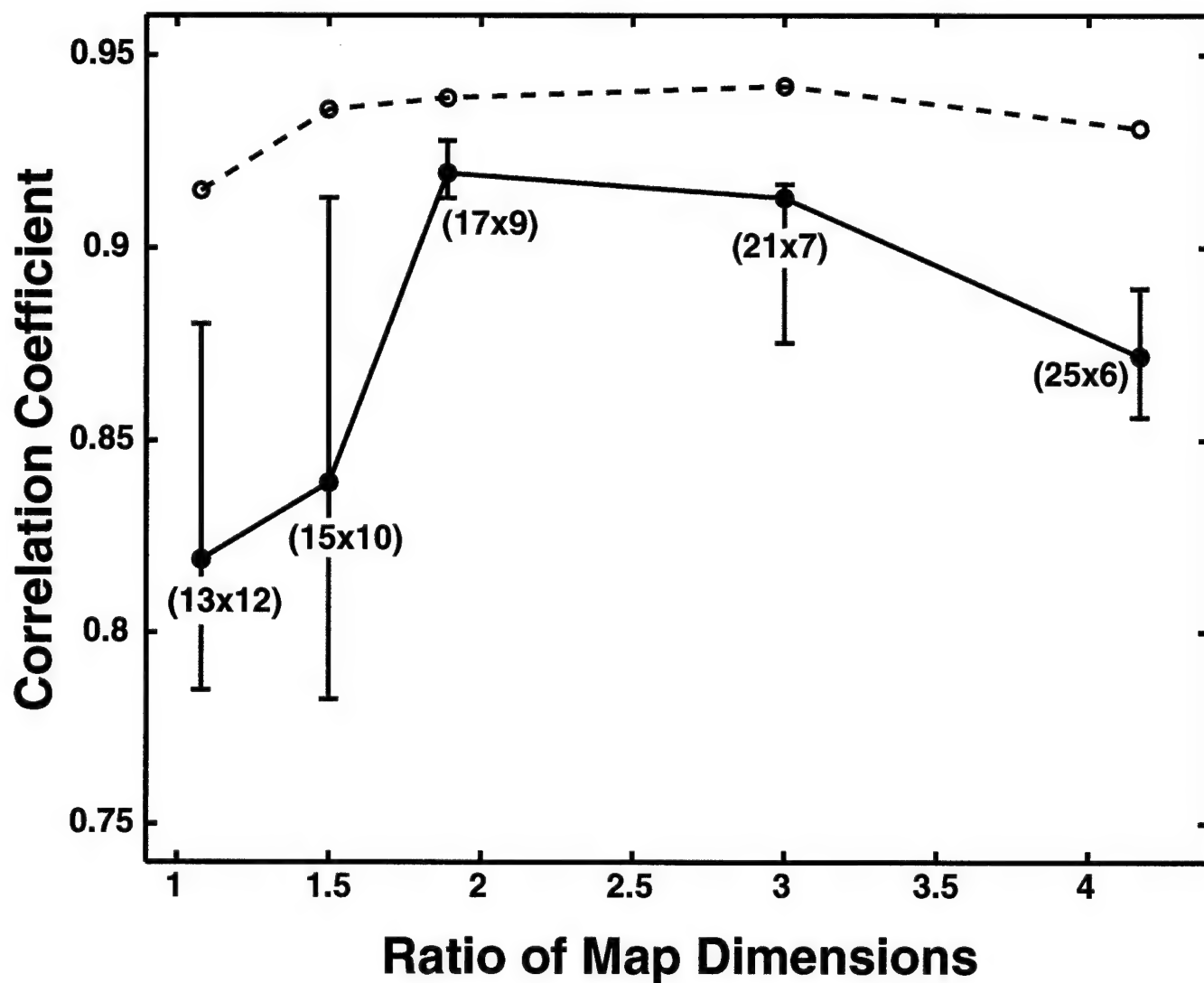
## **Structure/Function Correlation**



**Expanded Mechanism of Action Data Set**

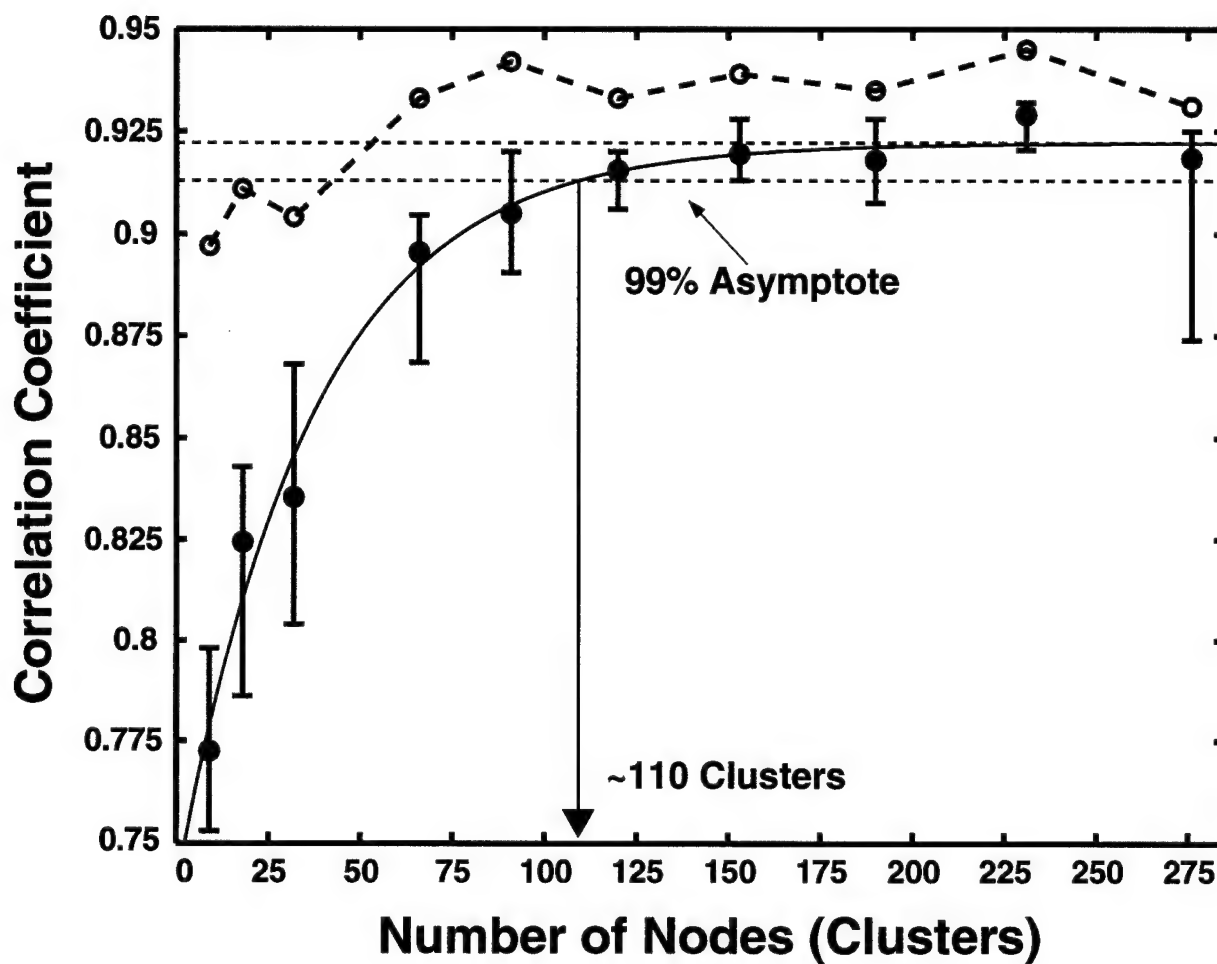
**Fig 3.**

**Structure/Function Correlation**  
**vs. Map Dimensions**



**Fig. 4**

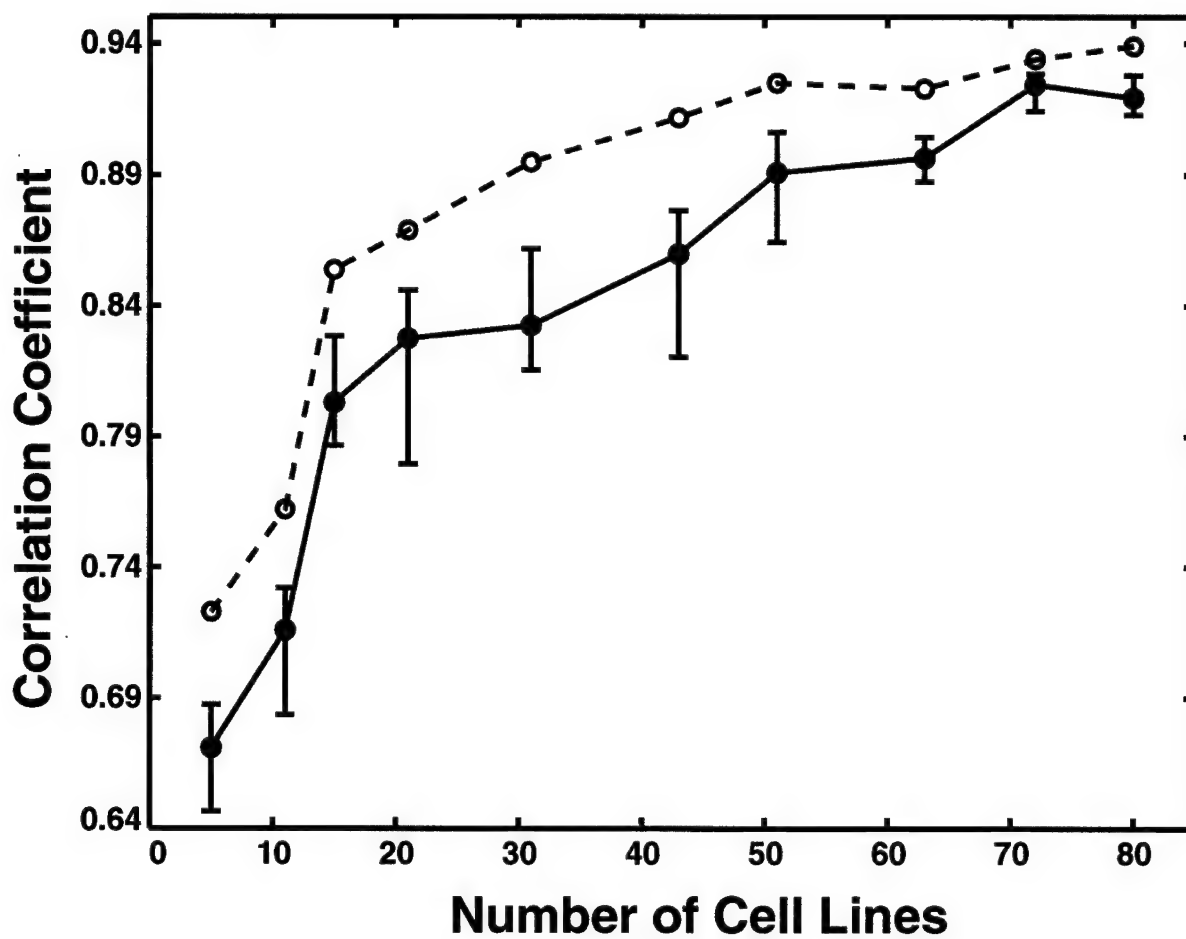
**Structure/Function Correlation**  
**vs. Number of Clusters**





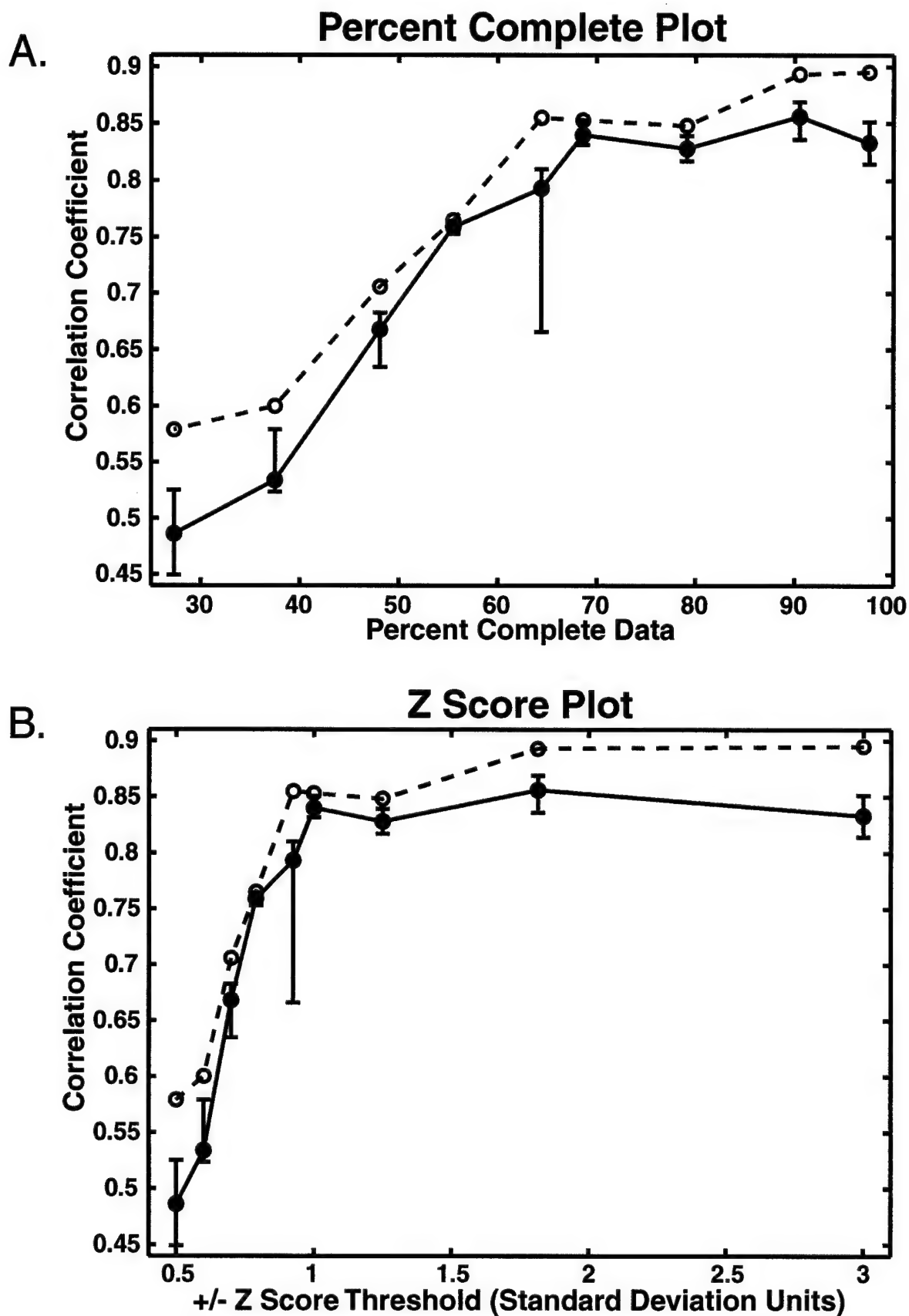
**Fig. 5**

**Structure/Function Correlation**  
**vs. Number Cell Lines**



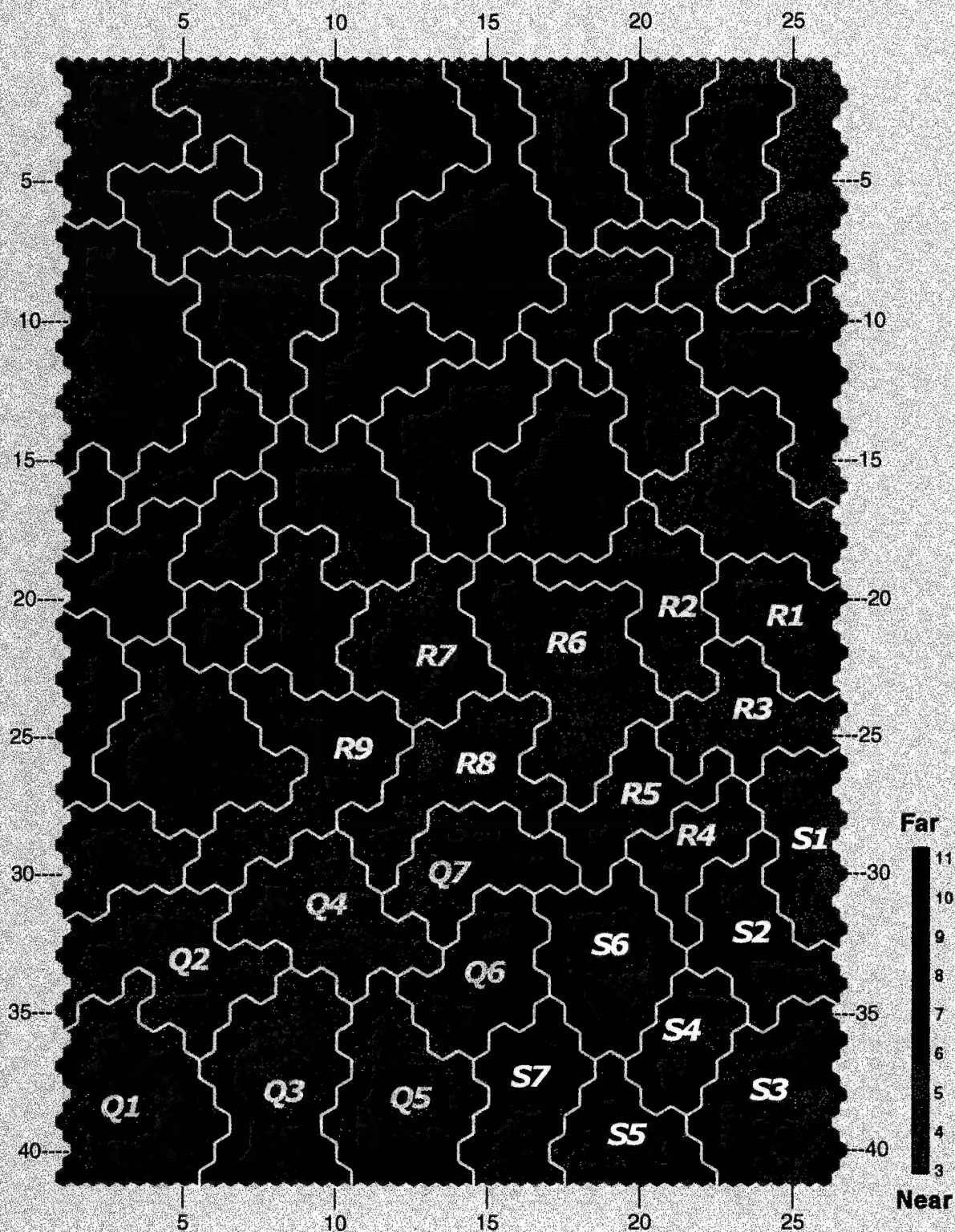
**Fig 6.**

**Structure/Function Correlation**  
**vs. Degradation of Data Set**



**Fig. 7**

## **Complete DTP SOM Map**

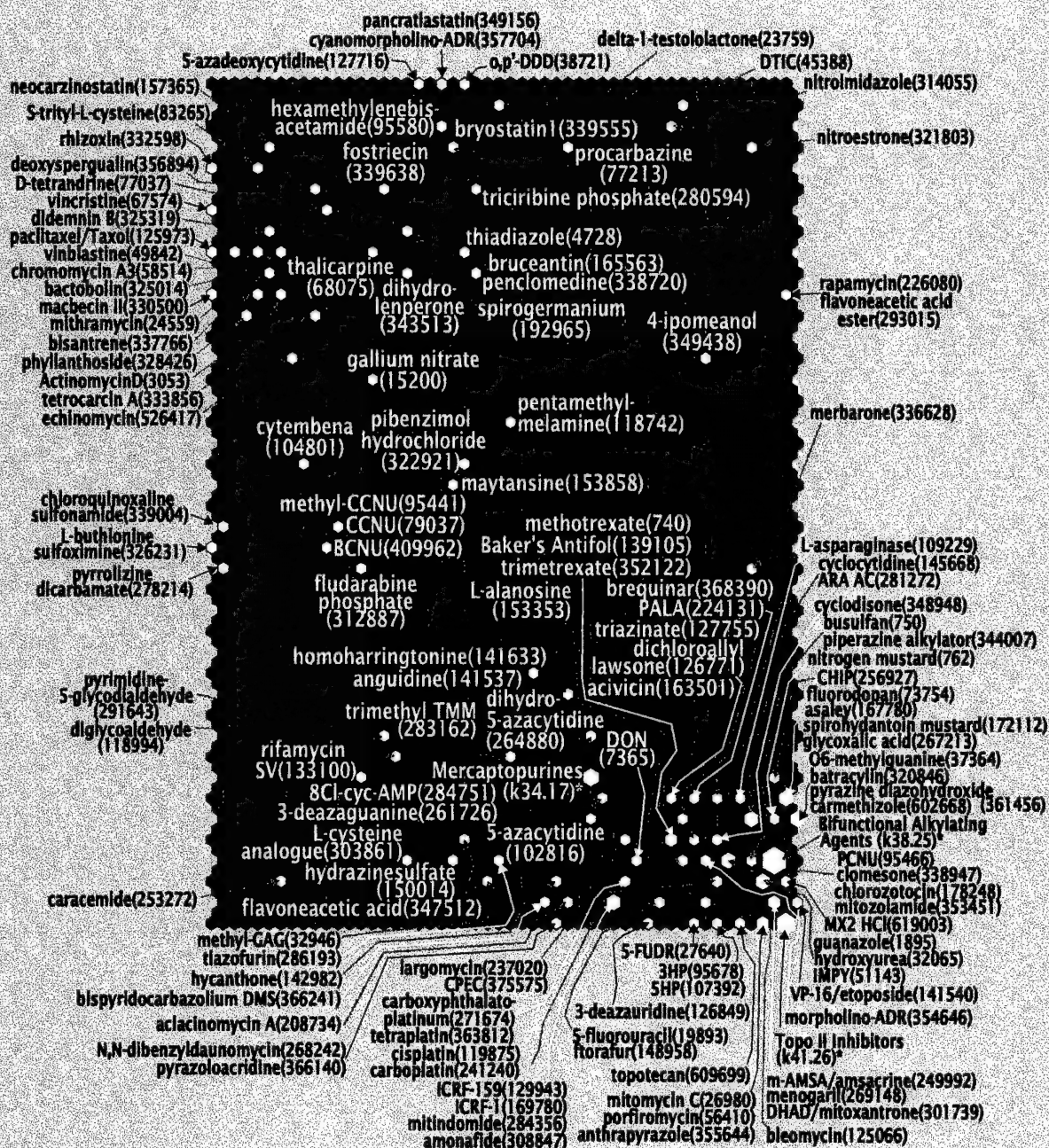


**Map Size 41x26**  
**(Aug 1999 Data Set, 20k Compounds)**



Fig. 8A

# Standard Agents

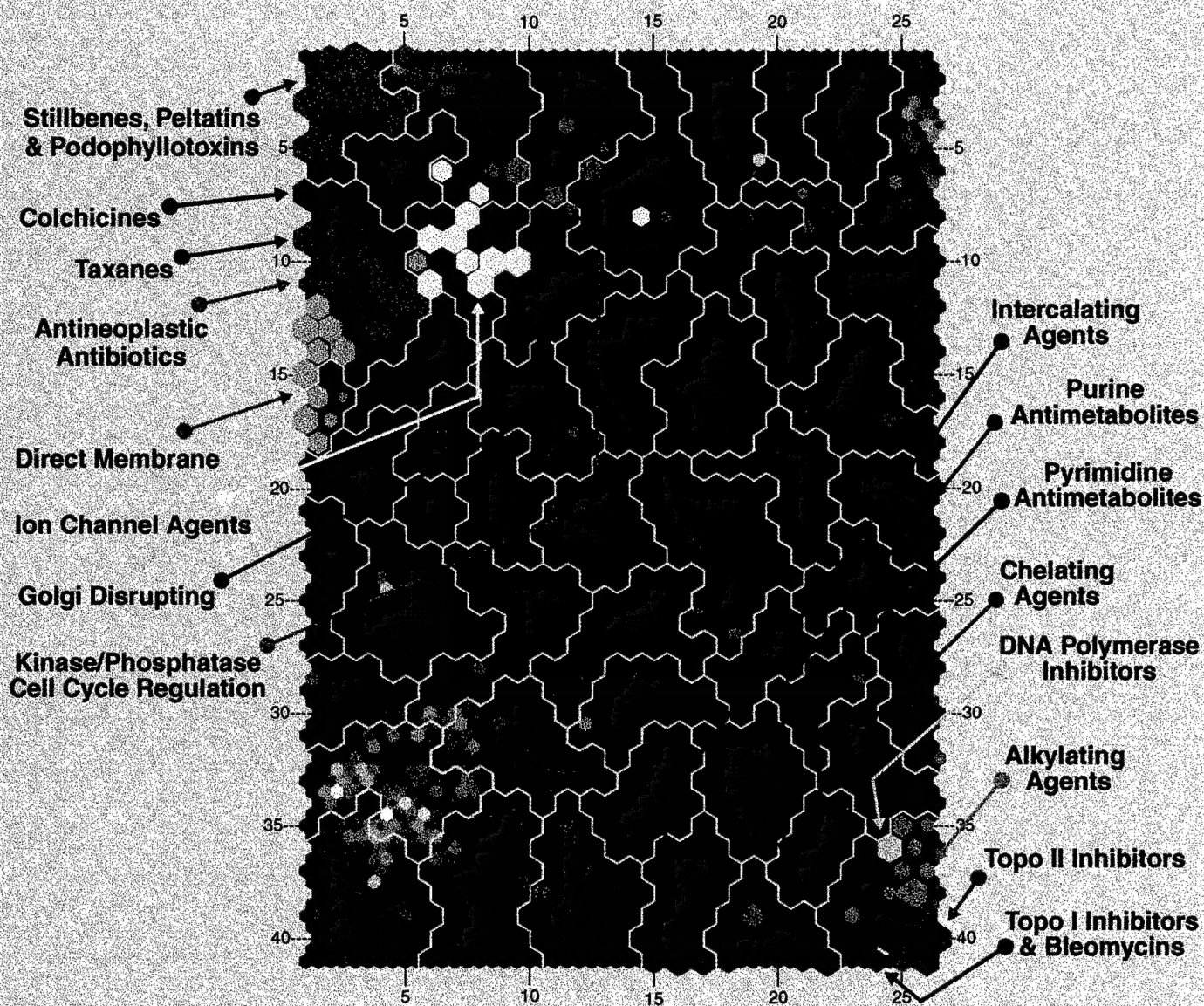


Complete Compound Map (Aug 1999 Data Set)



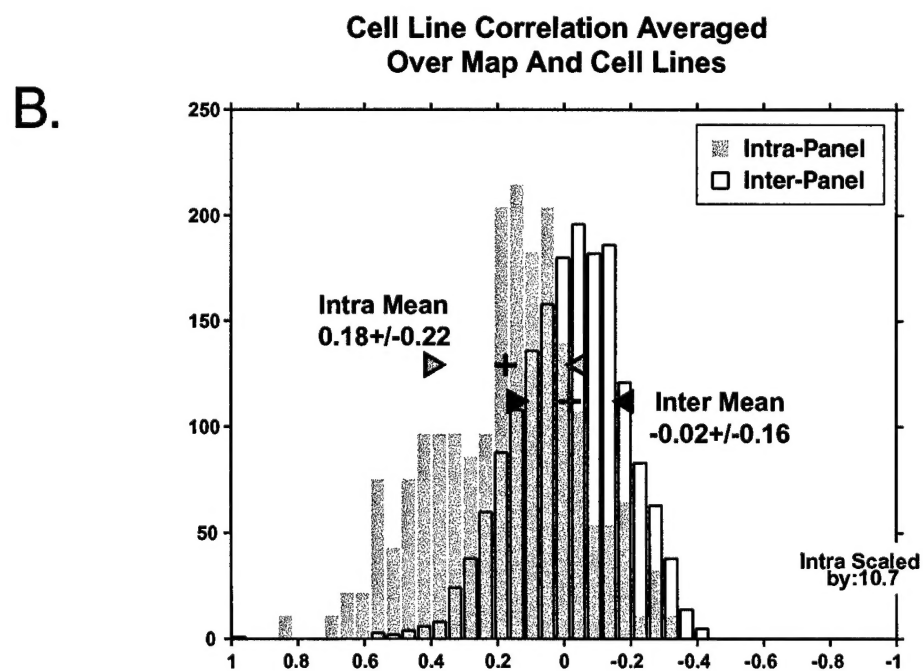
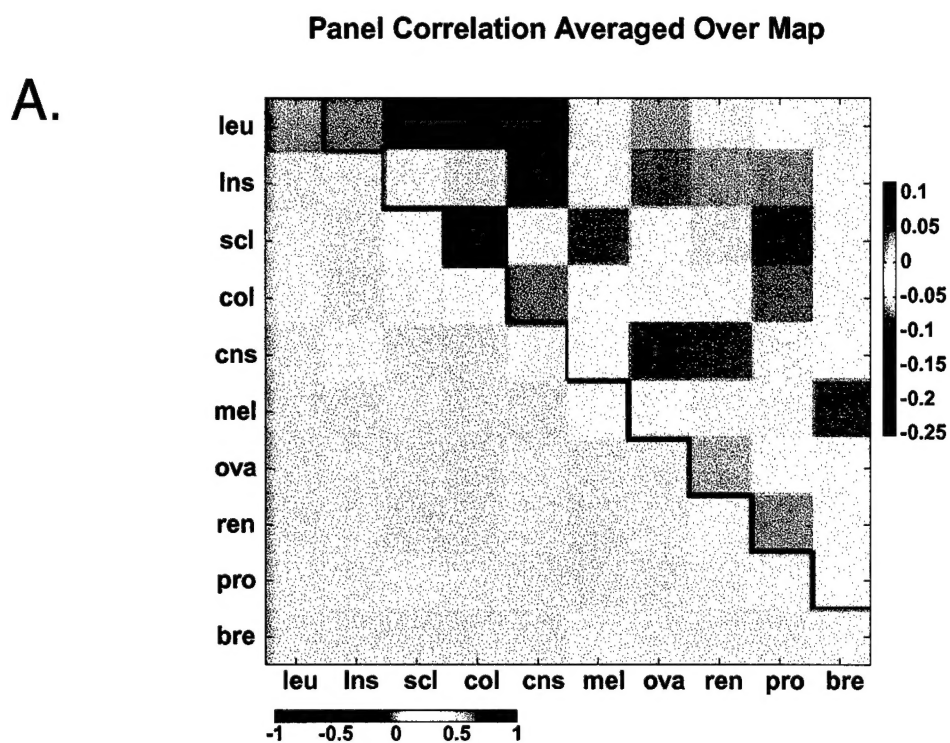
**Fig. 8B**

## Molecular Activity Classes



**Complete Compound Map (Aug 1999 Data Set)**

**Fig 9.**



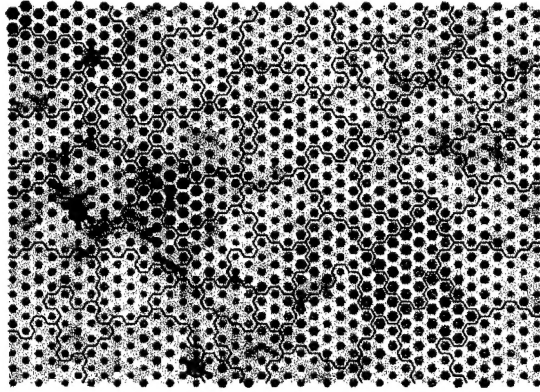


**Fig. 10**

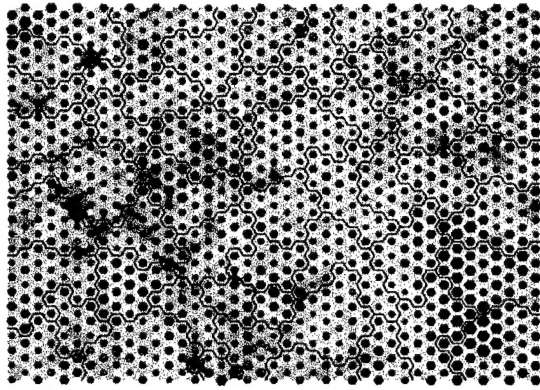
**Cell Panel Sensitivity**

◆ Sensitive  
● Insensitive

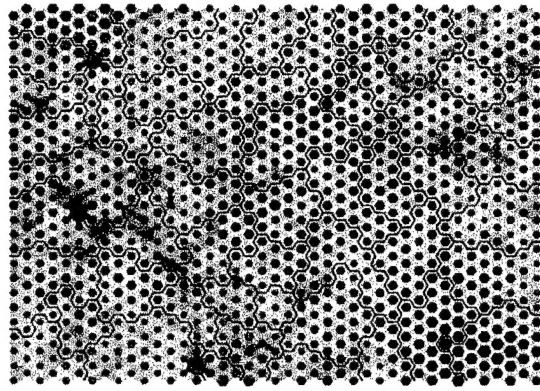
**Leu**



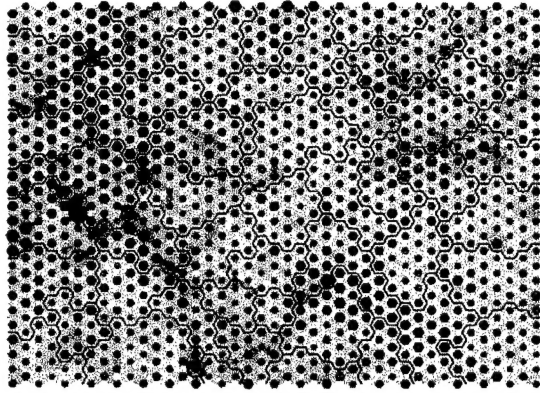
**Lns**



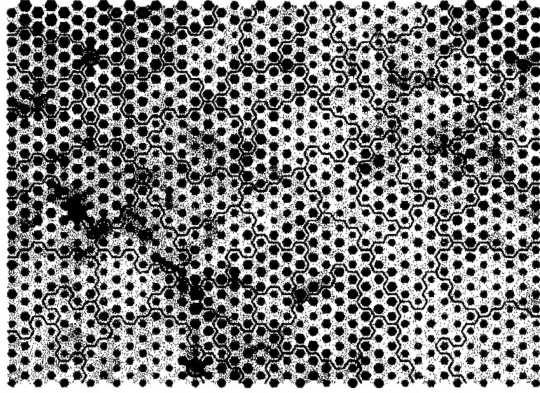
**Cns**



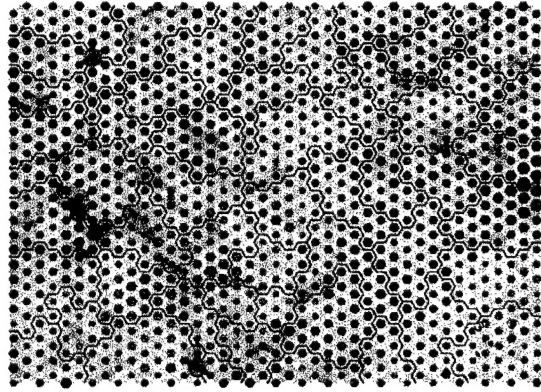
**Sci**



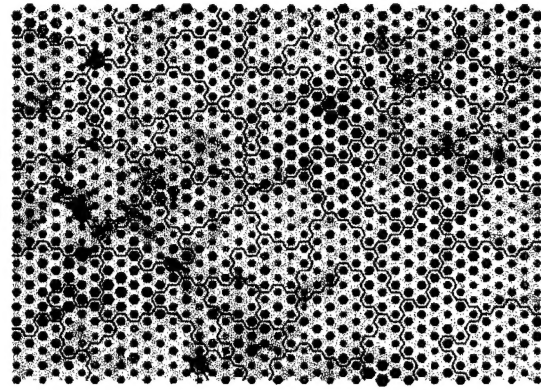
**Col**



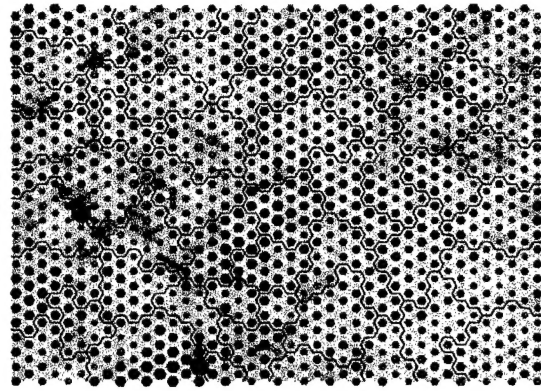
**Mel**



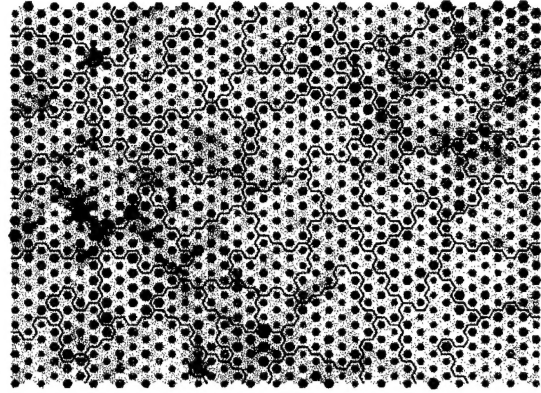
**Ova**



**Ren**



**Pro**



**Bre**

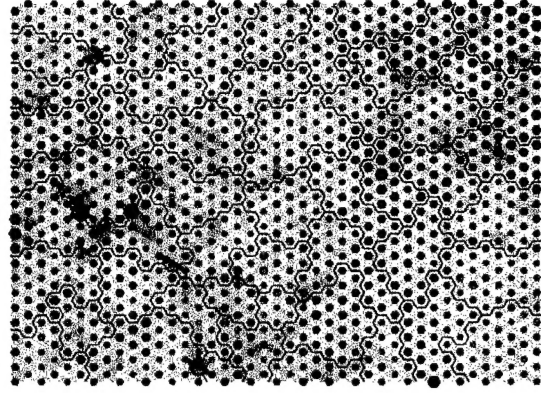
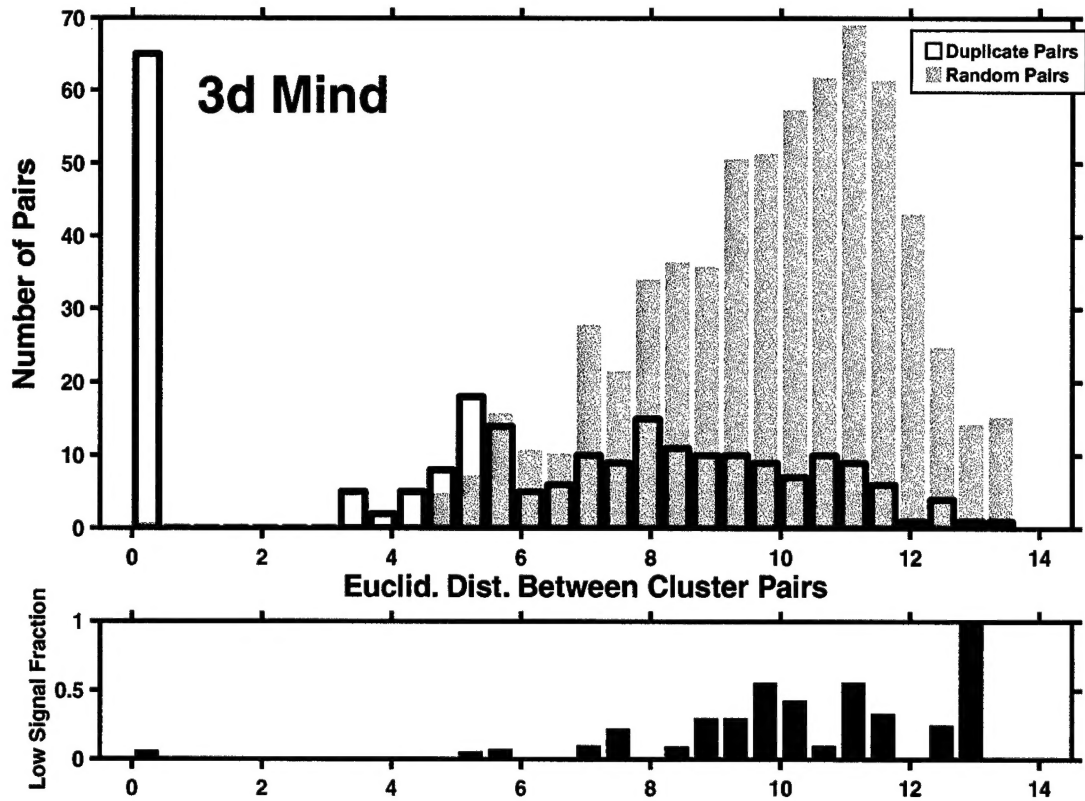


Fig. 11

A.



B.

